Sparse recovery by reduced variance stochastic approximation

Anatoli B. Juditsky, Andrei Kulunchakov, Hlib Tsyntseus

Sparse recovery by reduced variance stochastic approximation

Anatoli Juditsky[*,1] Andrei Kulunchakov[2] Hlib Tsyntseus[3,1]

March 31, 2022

Abstract

In this paper, we discuss application of iterative Stochastic Optimization routines to the problem of sparse signal recovery from noisy observation. Using Stochastic Mirror Descent algorithm as a building block, we develop a multistage procedure for recovery of sparse solutions to Stochastic Optimization problem under assumption of smoothness and quadratic minoration on the expected objective. An interesting feature of the proposed algorithm is linear convergence of the approximate solution during the preliminary phase of the routine when the component of stochastic error in the gradient observation which is due to bad initial approximation of the optimal solution is larger than the "ideal" asymptotic error component owing to observation noise "at the optimal solution." We also show how one can straightforwardly enhance reliability of the corresponding solution by using Median-of-Means like techniques.

We illustrate the performance of the proposed algorithms in application to classical problems of recovery of sparse and low rank signals in the generalized linear regression framework. We show, under rather weak assumption on the regressor and noise distributions, how they lead to parameter estimates which obey (up to factors which are logarithmic in problem dimension and confidence level) the best known to us accuracy bounds.

Keywords: sparse recovery, stochastic approximation, robust estimation

2000 Math Subject Classification: 62G08, 62G35, 62J07, 90C15

1 Introduction

In this paper, we consider the Stochastic Optimization problem of the form

$$g_* = \min_{x \in X} \big\{ g(x) = \mathbf{E}\{G(x,\omega)\} \big\} \tag{1}$$

where X is a given convex and closed subset of a Euclidean space E, $G: X \times \Omega \to \mathbf{R}$ is a smooth convex mapping, and $\mathbf{E}$ stands for the expectation with respect to unknown distribution of $\omega \in \Omega$ (we assume that the corresponding expectation exists for every $x \in X$). As it is usual in this situation, we suppose that we have access to a stochastic "oracle" supplying "randomized" information about g; we assume that the problem is solvable with the optimal solution x_* which is *sparse* (we consider a general notion of sparsity structure of x_* as defined in Section 2.1 which comprises "usual" sparsity, group sparsity, and low rank matrix structures as basic examples).

[*]LJK, Université Grenoble Alpes, 700 Avenue Centrale, 38401 Domaine Universitaire de Saint-Martin-d'Hères, France, anatoli.juditsky@univ-grenoble-alpes.fr

[1]Research of this author was supported by MIAI @ Grenoble Alpes (ANR-19-P3IA-0003).

[2]Université Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, 38000, France, andrei.kulunchakov@inria.fr

[3]LJK, Université Grenoble Alpes, 700 Avenue Centrale, 38401 Domaine Universitaire de Saint-Martin-d'Hères, France, hlib.tsyntseus@univ-grenoble-alpes.fr

Our interest in (1) is clearly motivated by statistical applications. Recently, different techniques of estimation and selection under sparsity and low rank constraints gained a lot of attention, in particular, in relation with the sparse linear regression problem in which unknown s-sparse (i.e., with at most s nonvanishing components) vector $x_* \in \mathbf{R}^n$ of regression coefficients is to be recovered from the linear noisy observation

$$\eta = \Phi^T x_* + \sigma\xi, \tag{2}$$

where $\Phi \in \mathbf{R}^{n\times N}$ is the regression matrix, and $\xi \in \mathbf{R}^N$ is zero-mean noise with unit covariance matrix; we are typically interested in the situation where the problem dimension is large, i.e. when $n \gg N$. Note that the problem of sparse recovery from observation (2) with random regressors (columns of the regression matrix Φ) ϕ_i, $i = 1, ..., N$ can be cast as Stochastic Optimization. For instance, assuming that regressors ϕ_i and noises ξ_i, $i = 1, ..., N$, are identically distributed, we may consider Stochastic Optimization problem

$$\min_{x\in X}\left\{g(x) = \tfrac{1}{2}\mathbf{E}\{(\eta_1 - \phi_1^T x)^2\}\right\} \tag{3}$$

over s-sparse $x \in X$. There are essentially two approaches to solving (3). Note that observations η_i and ϕ_i provide us with unbiased estimates $G(x, \omega_i = [\phi_i, \eta_i]) = \frac{1}{2}\|\eta_i - \phi_i^T x\|_2^2$ of the problem objective $g(x)$. Therefore, one can build a Sample Average Approximation (SAA)

$$\widehat{g}(x) = \frac{1}{N}\sum_{i=1}^N G(x, \omega_i) = \tfrac{1}{2N}\|\eta - \Phi^T x\|_2^2$$

of the objective $g(x)$ of (3) and then solve the resulting Least Squares problem by a deterministic optimization routine. A now standard approach to enhancing the sparsity of solutions is to use iterative thresholding [7, 22, 26, 41]. When applied to the linear regression problem (3), this technique amounts to using a gradient descent to minimize the Least Squares objective $\widehat{g}$ in combination with thresholding of approximate solutions to enforce sparsity. Another approach which refers to ℓ_1- and nuclear norm minimization allows to reduce problems of sparse or low rank recovery to convex optimization. In particular, sparse recovery by Lasso and Dantzig Selector has been extensively studied in the statistical literature [3, 12, 13, 14, 15, 16, 17, 19, 21, 28, 33, 44, 52, 54, 57], among others). For instance, the celebrated Lasso estimate $\widehat{x}_{N,\mathrm{lasso}}$ in the sparse linear regression problem is a solution to the ℓ_1-penalized Least Squares problem

$$\widehat{x}_{N,\mathrm{lasso}} \in \operatorname*{Argmin}_x \left\{\tfrac{1}{2N}\|\eta - \Phi^T x\|_2^2 + \lambda\|x\|_1\right\} \tag{4}$$

where $\lambda \geq 0$ is the algorithm parameter. Several conditions which ensure recovery with "small error" of any sparse or low rank signal using ℓ_1- and nuclear norm minimization are proposed. In particular, recovery of *any* s-sparse (i.e., with at most s nonvanishing components) vector x_* is possible with "small error" if the empirical regressor covariance matrix $\widehat{\Sigma} = \frac{1}{N}\Phi\Phi^T$ verifies a certain restricted conditioning assumption, e.g., Restricted Eigenvalue (RE) [3] or Compatibility condition [57]. The latter conditions very roughly mean that for all vectors z which are "approximately sparse," i.e., which are close to vectors with only s nonvanishing entries, $\|\widehat{\Sigma}z\|_2 \geq \lambda\|z\|_2$. The good news is that although these conditions are typically difficult to verify for individual matrices Φ, they are satisfied for several families of random matrices, such as Rademacher (with independent random ± 1 entries) and Gaussian matrices, matrices uniformly sampled from Fourier or Hadamard bases of $\mathbf{R}^n$, etc. For instance, when columns ϕ_i of Φ are sampled independently from normal distribution $\phi_i \sim \mathcal{N}(0, \Sigma)$ with covariance matrix Σ with

bounded diagonal elements which satisfies $\kappa_\Sigma I \preceq \Sigma$ (here I is the $n \times n$-identity matrix),[1] $\kappa_\Sigma > 0$, RE condition holds with high probability for s as large as $O\left(\frac{N\kappa_\Sigma}{\ln[n]}\right)$ [52].[2]

The Restricted Strong Convexity (RSC) condition, analogous to the RE or Compatibility condition also ensure that iterative thresholding procedures converge linearly to an approximate solution with accuracy which is similar to that of Lasso or Dantzig Selector estimation [22, 41] in this case.

Another approach to solving (1) which refers to Stochastic Approximation (SA) may be used whenever there is a "stochastic oracle" providing an unbiased stochastic observation of the gradient ∇g of the objective g of (1). For instance, note that the observable quantity $\nabla G(x, \omega_i) = \phi_i(\phi_i^T x - \eta_i)$ is an unbiased estimate of the gradient $\nabla g(x)$ of the objective of (3), and so an iterative algorithm of Stochastic Approximation type can be used to build approximate solutions to (3). In particular, different versions of Stochastic Approximation procedure were applied to solve (3) under ℓ_1 and sparsity constraint. Recall, that we are interested in high-dimensional problems, we are looking for bounds for recovery error which are "essentially independent" (logarithmic, at most) in problem dimension n. This requirement rules out the use of standard "Euclidean" Stochastic Approximation. Indeed, typical bounds for the expected inaccuracy $\mathbf{E}\{g(\widehat{x}_N)\} - g_*$ of Stochastic Approximation contains the term proportional to $\sigma^2\mathbf{E}\{\|\phi_1\|_2^2\}$ and thus proportional to n in the case of "dense" regressors with $\mathbf{E}\{\|\phi_1\|_2^2\} = O(n)$. Therefore, unless regressors ϕ are sparse (or possess a special structure, e.g., when ϕ_i are low rank matrices in the case of low rank matrix recovery), standard Stochastic Approximation leads to accuracy bounds for sparse recovery which are proportional to dimension n of the parameter vector [50]. In other words, our application calls for non-Euclidean Stochastic Approximation procedures, such as Stochastic Mirror Descent algorithm [46].

In particular, [55, 56] study the properties of Stochastic Mirror Descent algorithm under sub-Gaussian noise assumption and show that approximate solution $\widehat{x}_N$ after N iterations of the method attains the bound $g(\widehat{x}_N) - g_* = O\left(\sigma\sqrt{s\ln(n)/N}\right)$, often referred to as "slow rate" of sparse recovery. In order to improve the error estimates of Stochastic Approximation one may use multistage algorithm under strong or uniform convexity assumption [24, 29, 30]. However, such assumptions do not hold in the problems such as sparse linear regression problem,[3] where they are replaced by Restricted Strong Convexity conditions. For instance, the authors of [2, 23] develop a multistage procedure targeted at sparse recovery stochastic optimization problem (1) based on SMD algorithm of [31, 47] under bounded regressor and sub-Gaussian noise assumption. They show, for instance, that when applied to the sparse linear regression, the ℓ_2-error $\|\widehat{x}_N - x_*\|_2$ of the approximate solution $\widehat{x}_N$ after N iterations of the proposed routine converges at the rate $O\left(\frac{\sigma}{\kappa_\Sigma}\sqrt{\frac{s\ln n}{N}}\right)$ with high probability. While this "asymptotic" rate coincides with the best

[1] Here and in the sequel, we use notation $A \preceq B$ for $n \times n$ symmetric matrices A and B such that $B - A \succeq 0$, i.e. $B - A$ is positive semidefinite.

[2] The reader acquainted with the compressive sensing theory will notice that the setting of the ℓ_1-recovery problem considered in this paper is different from the s"tandard setting," but is rather similar in spirit to that in [1, 5, 8, 13, 17]. Although, unlike [1, 5, 8] we do not assume any special structure of x_* apart from its sparsity, we suppose random regressors to be independent of x_*, while in the "standard setting" one allows for the "worst case x_*" which may depend on the particular realization of the matrix of regressors. Nevertheless, we do not know any result stating that a recovery in the present setting is possible under "essentially less restrictive" assumptions than those for the "standard" ℓ_1 recovery.

[3] More generally, strong convexity of the objective associated with smoothness is a feature of the Euclidean setup. For instance, the conditioning of a smooth objective (the ratio of the Lipschitz constant of the gradient to the constant of strong convexity) when measured with respect to the ℓ_1-norm cannot be less than n (the problem dimension) [30].

rate attainable by known to us algorithms for solving (3) the algorithm in [2, 23] requires at least $\frac{s^2\ln[n]}{\kappa_\Sigma^2}$ SMD iterations per stage, implying that the method in question can be used only if the number of nonvanishing entries in the parameter vector is $O\left(\kappa_\Sigma\sqrt{\frac{N}{\ln n}}\right)$[4] (recall that the corresponding limit is $O\left(\frac{N\kappa_\Sigma}{\ln[n]}\right)$ for Lasso [52] and iterative thresholding procedures [22, 41]).

Our goal in the present paper is to provide a refined analysis of Stochastic Approximation algorithms for computing sparse solutions to (1) exploiting a variance reduction scheme utilizing in a special way smoothness of the problem objective.[5] It allows to build a new accelerated multistage Stochastic Approximation algorithm. To give a flavor of the results we present below, we summarize the properties of the proposed procedure—Stochastic Mirror Descent for Sparse Recovery (SMD-SR)—in the case of stochastic optimization problem (3) associated with sparse linear regression estimation problem. Let us assume that regressors ϕ_i are a.s. bounded, i.e., $\|\phi_i\|_\infty = O(1)$, the covariance matrix $\Sigma = \mathbf{E}\{\phi_1\phi_1^T\}$ of regressors satisfies $\Sigma \succeq \kappa_\Sigma I$; we suppose that the noises $\sigma\xi_i$ are zero-mean with $\mathbf{E}\{\xi_i^2\} \le 1$, and that we are given $R < \infty$ and $x_0 \in \mathbf{R}^n$ such that $\mathbf{E}\{\|x_0 - x_*\|_1^2\} \le R^2$.

- The SMD-SR algorithm is organized in stages. On the k-th stage of the method we run N_k iterations of the Stochastic Mirror Descent recursion and then "sparsify" the obtained approximate solution by zeroing out all but s entries of largest amplitudes.

- Stages of the algorithm are organized into two groups (phases). At the first (preliminary) phase we perform a fixed number $N_k = O\left(\frac{s\ln n}{\kappa_\Sigma}\right)$ of SMD iterations per stage to guarantee that the expected quadratic error $\mathbf{E}\{\|\widehat{y}_k - x_*\|_1^2\}$ of the sparse approximate solution $\widehat{y}_k$ of the k-th stage is smaller than the expected error $\mathbf{E}\{\|\widehat{y}_{k-1} - x_*\|_1^2\}$ of the previous stage solution y_{k-1} by a fixed factor. Thus, the error of the approximate solution after (total) N iterations decreases linearly with the exponent proportional to $\frac{\kappa_\Sigma}{s\ln n}$. When the expected quadratic error becomes $O\left(\frac{\sigma^2 s^2}{\kappa_\Sigma}\right)$, we pass to the second (asymptotic) phase of the method.

- During the stages of the asymptotic phase, the number of iterations per stage grows as $N_k = 2^k N_0$ where k is the stage index, and the expected quadratic error decreases as $O\left(\frac{\sigma^2 s^2\ln n}{\kappa_\Sigma^2 N}\right)$ where N is total iteration count.

It may appear surprising that a stochastic algorithm converges linearly during the preliminary phase, when the component of the error due to the observation noise is small (for instance, it converges linearly in the "noiseless" case, cf. [50]) eliminating fast the initial error; its rate of convergence is similar to that of the deterministic gradient descent algorithm, when "full gradient observation" $\nabla g(x)$ is available. On the other hand, in the asymptotic regime, the procedure attains the rate which is equivalent to the best known rates in this setting, and under the model assumptions which are close to the weakest known today [41, 52].

The paper is organized as follows. The analysis of the SMD-SR in the general setting is in Section 2. We define the general problem setting and introduce key notions used in the paper in Section 2.1. Then in Section 2.3 we reveal the multistage algorithm and study its basic properties. Next, in Section 2.4 we show how sub-Gaussian confidence bounds for the error of

[4] That being said, [2], for instance, deals with *nonsmooth* stochastic optimization, so the scope of corresponding algorithms is much larger than the framework of smooth problems considered in this paper.

[5] In hindsight, the underlying idea can be seen as a generalization of the variance reduction device in [4].

approximate solutions can be obtained using an adopted analog of Median-of-Means approach. Finally, in Section 3 we discuss the properties of the method and conditions in which it leads to "small error" solution when applied to sparse linear regression and low rank linear matrix recovery problems.

2 Sparse solutions to stochastic optimization problem

2.1 Problem statement

Let E be a finite-dimensional real vector (Euclidean) space. Consider a Stochastic Optimization problem

$$\min_{x\in X}\left[\mathbf{E}\{G(x,\omega)\}\right] \tag{5}$$

where $X \subset E$ is a convex set with nonempty interior (a solid), ω is a random variable on a probability space Ω with distribution P, and $G:\ X\times\Omega\to\mathbf{R}$. We suppose that the expected objective

$$g(x)=\mathbf{E}\{G(x,\omega)\}$$

is finite for all $x\in X$ and is convex and differentiable on X. Let $\|\cdot\|$ be a norm on E, and let $\|\cdot\|_*$ be the conjugate norm, i.e.,

$$\|s\|_*=\max_x\{\langle s,x\rangle:\ \|x\|\le 1\},\quad s\in E.$$

We suppose that gradient $\nabla g(\cdot)$ of $g(\cdot)$ is Lipschitz-continuous on X:

$$\|\nabla g(x')-\nabla g(x)\|_*\le\mathcal{L}\|x-x'\|,\qquad \forall\, x,x'\in X, \tag{6}$$

that the problem is solvable with optimal value $g_*=\min_{x\in X}g(x)$. Furthermore, we suppose that the optimal solution x_* to the problem is unique, and that $g(\cdot)$ satisfies *quadratic growth condition* on X with respect to the Euclidean norm $\|\cdot\|_2$ [43], i.e., for all $x\in X$

$$g(x)-g_*\ge\tfrac{1}{2}\underline{\kappa}\|x-x_*\|_2^2 \tag{7}$$

where $\|\cdot\|_2$ is the Euclidean norm: $\|z\|_2=\langle z,z\rangle^{1/2}$. In what follows, we assume that we have at our disposal a *stochastic* (gray box) *oracle*—a device which can generate $\omega\sim P$ and compute, for any $x\in X$ a random unbiased estimation of $\nabla g(x)$. From now on we make the following assumption about the structure of the gradient observation:

Assumption [S1]. *$G(\cdot,\omega)$ is differentiable on X for almost all $\omega\in\Omega$, and*[6]

$$\mathbf{E}\{\nabla G(x,\omega)\}=\nabla g(x)\quad \textit{and}\quad \mathbf{E}\{\|\underbrace{\nabla G(x,\omega)-\nabla g(x)}_{=:\zeta(x,\omega)}\|_*^2\}\le\varsigma^2(x),\quad\forall\, x\in X.$$

Furthermore, there are $1\le\varkappa,\varkappa'<\infty$ and $\mathcal{L}\le\nu<\infty$ such that the bound holds:

$$\varsigma^2(x)\le\varkappa\nu[g(x)-g_*-\langle\nabla g(x_*),x-x_*\rangle]+\varkappa'\underbrace{\mathbf{E}\{\|\zeta(x_*,\omega)\|_*^2\}}_{=:\varsigma_*^2}. \tag{8}$$

[6] In what follows $\nabla G(\cdot,\omega)$ replaces notation $\nabla_x G(\cdot,\omega)$ for the gradient of G w.r.t. the first argument.

Remarks. Assumption **S1** and, in particular, bound (8) are essential to the subsequent developments and certainly merit some comments. We postpone the corresponding discussion to Section 3 where we present several examples of observation models in which this assumption naturally holds. For now, let us consider a simple example of the Stochastic Optimization problem (3) arising in sparse regression estimation where regressors ϕ_i are a.s. bounded, i.e., $\|\phi_i\|_\infty \le r < \infty$ with identity covariance matrix $\mathbf{E}\{\phi_1\phi_1^T\} = I$, and noises $\sigma\xi_i$ are zero-mean with "small" variance. In the situation in question, the error $\zeta(x,\omega) = \nabla G(x,\omega) - \nabla g(x)$, $\omega = [\phi, \xi]$, of the stochastic oracle can be decomposed as in

$$\zeta(x,\omega) = \underbrace{[\phi\phi^T - I](x - x_*)}_{=:\zeta_1(x,\omega)} + \underbrace{\sigma\xi\phi}_{=:\zeta_2(\omega)}.$$

Note that the "variance" $\varsigma_1^2(x)$ of the first component satisfies

$$\varsigma_1^2(x) = \mathbf{E}\{\|\zeta_1(x,\omega)\|_\infty^2\} \le 2(r^2+1)\|x - x_*\|_2^2 \le 4(r^2+1)(g(x) - g_*),$$

while the "variance" ς_2^2 of the second,

$$\varsigma_2^2 = \mathbf{E}\{\|\zeta_2(\omega)\|_\infty^2\} = \sigma^2\mathbf{E}\{\|\phi\|_\infty^2\} \le \sigma^2 r^2,$$

does not depend on x. As a result, the bound

$$\varsigma^2(x) = \mathbf{E}\{\|\zeta(x,\omega)\|_\infty^2\} \le 4(r^2+1)\|x - x_*\|_2^2 + 2\sigma^2 r^2$$

implies that in this case the stochastic gradient observation $\nabla G(x,\omega)$ satisfies Assumption **S1** with $\varsigma_*^2 = \sigma^2 r^2$, $\nu = r^2 + 1$, $\kappa = 8$ and $\kappa' = 2$.

More generally, relation (8) is rather characteristic to the case of smooth stochastic observation. Indeed, let us consider the situation where the stochastic gradient $G(\cdot,\omega)$ itself is Lipschitz-continuous on X with a.s. bounded Lipschitz constant $\mathcal{L}(\omega)$ with respect to the norm $\|\cdot\|$, $\mathcal{L}(\omega) \le \nu$. In this case we have

$$\begin{aligned}\varsigma^2(x) &= \mathbf{E}\big\{\|\nabla G(x,\omega) - \nabla g(x)\|_*^2\big\} \le \Big(\mathbf{E}\big\{\|\nabla G(x,\omega) - \nabla G(x_*,\omega)\|_*^2\big\}^{1/2} \\ &\quad + \|\nabla g(x) - \nabla g(x_*)\|_* + \mathbf{E}\big\{\|\nabla G(x_*,\omega) - \nabla g(x_*)\|_*^2\big\}^{1/2}\Big)^2.\end{aligned}$$

However, due to the Lipschitz continuity of $\nabla G(\cdot,\omega)$

$$G(x,\omega) - G(x_*,\omega) \ge \langle \nabla G(x_*,\omega), x - x_*\rangle + (2\nu)^{-1}\|\nabla G(x,\omega) - \nabla G(x_*,\omega)\|_*^2,$$

implying that

$$\begin{aligned}\varsigma^2(x) &\le \Big([2\nu\mathbf{E}\{G(x,\omega) - G(x_*,\omega) - \langle \nabla G(x_*,\omega), x - x_*\rangle\}]^{1/2} \\ &\quad + [2\nu(g(x) - g(x_*) - \langle \nabla g(x_*), x - x_*\rangle)]^{1/2} + \varsigma_*\Big)^2 \\ &\le 16\nu[g(x) - g_* - \langle \nabla g(x_*), x - x_*\rangle] + 2\varsigma_*^2.\end{aligned}$$

Sparsity structure. In what follows we assume to be given a *sparsity structure* [32] on E—a family $\mathcal{P}$ of projector mappings $P = P^2$ on E with associated nonnegative weights $\pi(P)$. For a nonnegative real s we set

$$\mathcal{P}_s = \{P \in \mathcal{P} : \pi(P) \le s\}.$$

Given $s \ge 0$ we call $x \in E$ *s-sparse* if there exists $P \in \mathcal{P}_s$ such that $Px = x$. We will make the following standing assumption.

Assumption [S2] *The optimal solution x_* to problem (5) is s-sparse.*

Furthermore, given $x \in X$ one can efficiently compute a "sparse approximation" of x—an optimal solution $x_s = \mathrm{sparse}(x)$ to the optimization problem

$$\min \|x - z\|_2 \ \ \textit{over } s\textit{-sparse } z \in X. \tag{9}$$

Moreover, for any s-sparse $z \in E$ the norm $\|\cdot\|$ satisfies $\|z\| \le \sqrt{s}\|z\|_2$.

In what follows we refer to x_s as "sparsification of x." We are mainly interested in the following "standard examples":

1. "Vanilla" sparsity: in this case $E = \mathbf{R}^n$ with the standard inner product, $\mathcal{P}$ is comprised of projectors on all coordinate subspaces of $\mathbf{R}^n$, $\pi(P) = \mathrm{rank}(P)$, and $\|\cdot\| = \|\cdot\|_1$.

   Assumption **S2** clearly holds, for instance, when X is orthosymmetric, e.g., a ball of ℓ_p-norm on $\mathbf{R}^n$, $1 \le p \le \infty$.

2. Group sparsity: $E = \mathbf{R}^n$, and we partition the set $\{1, ..., n\}$ of indices into K nonoverlapping subsets $I_1, ..., I_K$, so that to every $x \in \mathbf{R}^n$ we associate blocks x^k with corresponding indices in I_k, $k = 1, ..., K$. Now $\mathcal{P}$ is comprised of projectors $P = P_I$ onto subspaces $E_I = \{[x^1, ..., x^K] \in \mathbf{R}^n : x^k = 0 \, \forall k \notin I\}$ associated with subsets I of the index set $\{1, ..., K\}$. We set $\pi(P_I) = \mathrm{card} I$, and define $\|x\| = \sum_{k=1}^K \|x_k\|_2$—*block ℓ_1/ℓ_2-norm.*

   Same as above, Assumption **S2** holds in this case when X is "block-symmetric," for instance, is a ball of block norm $\|\cdot\|$.

3. Low rank sparsity structure: in this example $E = \mathbf{R}^{p \times q}$ with, for the sake of definiteness, $p \ge q$, and the Frobenius inner product. Here $\mathcal{P}$ is the set of mappings $P(x) = P_\ell x P_r$ where P_ℓ and P_r are, respectively, $q \times q$ and $p \times p$ orthoprojectors, and $\|\cdot\|$ is the nuclear norm $\|x\| = \sum_{i=1}^q \sigma_i(x)$ where $\sigma_1(x) \ge \sigma_2(x) \ge ... \ge \sigma_q(x)$ are singular values of x.

   In this case Assumption **S2** holds due to the Eckart–Young approximation theorem, it suffices that X is a ball of a Schatten norm $\|x\|_r = \left(\sum_{i=1}^q \sigma_i^r(x)\right)^{1/r}$, $1 \le r \le \infty$.

Our objective is to build approximate solutions $\widehat{x}_N$ to problem (5) utilizing N queries to the stochastic oracle. We quantify the performance of such solutions on the class $\mathcal{X} = \mathcal{X}(X, \mathcal{L}, ..., \mathcal{P}, s)$ of Sparse Stochastic Optimization problems (5) described in the beginning of this section satisfying Assumptions **S1** and **S2**, with domain X, by the following worst-case over $\mathcal{X}$ risk measures:

- *Recovery risks:* maximal over $\mathcal{X}$ expected squared error

  $$\mathrm{Risk}_{|\cdot|}(\widehat{x}|\mathcal{X}) = \sup_{\mathcal{X}} \mathbf{E}\{|\widehat{x} - x_*|^2\}^{1/2}$$

  where $|\cdot|$ stands for $\|\cdot\|_2$- or $\|\cdot\|$-norm, and ϵ-*risk* of recovery—the smallest maximal over $\mathcal{X}$ radius of $(1-\epsilon)$-confidence ball of norm $|\cdot|$ centered at $\widehat{x}$:

  $$\mathrm{Risk}_{|\cdot|,\epsilon}(\widehat{x}|\mathcal{X}) = \inf\left\{r : \sup_{\mathcal{X}} \mathrm{Prob}\{|\widehat{x} - x_*| \ge r\} \le \epsilon\right\}$$

- *Prediction risks:* maximal over $\mathcal{X}$ expected suboptimality

  $$\mathrm{Risk}_g(\widehat{x}|\mathcal{X}) = \sup_{\mathcal{X}} \mathbf{E}\{g(\widehat{x})\} - g_*,$$

  of $\widehat{x}$ and the smallest maximal over $\mathcal{X}$ $(1-\epsilon)$-confidence interval

  $$\mathrm{Risk}_{g,\epsilon}(\widehat{x}|\mathcal{X}) = \inf\left\{r : \sup_{\mathcal{X}} \mathrm{Prob}\{g(\widehat{x}) - g_* \ge r\} \le \epsilon\right\}. \tag{10}$$

In what follows, we use a generic notation c and C for absolute constants; notation $a \lesssim b$ means that the ratio a/b is bounded by an absolute constant.

2.2 Stochastic Mirror Descent algorithm

Notation and definitions. Let $\vartheta : E \to \mathbf{R}$ be a continuously differentiable convex function which is strongly convex with respect to the norm $\|\cdot\|$, i.e.,

$$\langle \nabla\vartheta(x) - \nabla\vartheta(x'), x - x' \rangle \geq \|x - x'\|^2, \quad \forall x, x' \in E.$$

From now on, w.l.o.g. we assume that $\vartheta(x) \geq \vartheta(0) = 0$. We say that Θ is the constant of quadratic growth of $\vartheta(\cdot)$ if

$$\forall x \in E \; \vartheta(x) \leq \Theta\|x\|^2.$$

Clearly, $\Theta \geq \frac{1}{2}$. If, in addition, Θ is "not too large," and for any $x \in X$, $a \in E$ and $\beta > 0$ a high accuracy solution to the minimization problem

$$\min_{z \in X}\{\langle a, z\rangle + \beta\vartheta(z - x)\}$$

can be easily computed, following [29, 30, 45, 49] we say that *distance-generating function (d.-g.f.)* ϑ is "prox-friendly." We present choices of prox-friendly d.-g.f.'s relative to the norm used in application sections.

We also utilize associated Bregman divergence

$$V_{x_0}(x, z) = \vartheta(z - x_0) - \vartheta(x - x_0) - \langle \nabla\vartheta(x - x_0), z - x\rangle, \quad \forall\, z, x, x_0 \in X.$$

For $Q \in \mathbf{R}^{p\times q}$ we denote

$$\|Q\|_\infty = \max_{ij} |[Q]_{ij}|;$$

for symmetric positive-definite $Q \in \mathbf{R}^{n\times n}$ and $x \in \mathbf{R}^n$ we denote

$$\|x\|_Q = \sqrt{x^T Q x}.$$

Stochastic Mirror Descent algorithm. For $x, x_0 \in X$, $u \in E$, and $\beta > 0$ consider the *proximal mapping*

$$\begin{array}{rcl} \mathrm{Prox}_\beta(u, x; x_0) & := & \mathrm{argmin}_{z\in X}\big\{\langle u, z\rangle + \beta V_{x_0}(x, z)\big\} \\ & = & \mathrm{argmin}_{z\in X}\big\{\langle u - \beta\langle \nabla\vartheta(x - x_0), z\rangle + \beta\vartheta(z - x_0)\big\}. \end{array} \tag{11}$$

For $i = 1, 2, \ldots,$ consider *Stochastic Mirror Descent* recursion, cf. [29, 36, 45],

$$x_i \;=\; \mathrm{Prox}_{\beta_{i-1}}(\nabla G(x_{i-1}, \omega_i), x_{i-1}; x_0), \quad x_0 \in X, \tag{12}$$

Here $\beta_i > 0$, $i = 0, 1, \ldots,$ is a stepsize parameter to be defined later, and $\omega_1, \omega_2, \ldots$ are independent identically distributed (i.i.d.) realizations of random variable ω, corresponding to the oracle queries at each step of the algorithm.

The approximate solution to problem (5) after N iterations is defined as weighted average

$$\widehat{x}_N = \left[\sum_{i=1}^N \beta_{i-1}^{-1}\right]^{-1} \sum_{i=1}^N \beta_{i-1}^{-1} x_i. \tag{13}$$

The next result describes some useful properties of the recursion (12).

Proposition 2.1 *Suppose that SMD algorithm is applied to problem (5) in the situation described in this section. We assume that Assumption* **S1** *holds and that initial condition $x_0 \in X$ is independent of ω_i, $i = 1, 2, ...$ and such that $\mathbf{E}\{\|x_0 - x_*\|^2\} \le R^2$; we use constant stepsizes*

$$\beta_i \equiv \beta \ge 2\varkappa\nu, \quad i = 1, 2, ..., m.$$

Then approximate solution $\widehat{x}_m = \frac{1}{m}\sum_{i=1}^m x_i$ after m steps of the algorithm satisfies

$$\mathbf{E}\{g(\widehat{x}_m)\} - g_* \le \frac{2R^2}{m}\left(\Theta\beta + \frac{\varkappa\nu^2}{2\beta}\right) + \frac{2\varkappa'\varsigma_*^2}{\beta}. \tag{14}$$

2.3 Multistage SMD algorithm

We assume to be given $R < \infty$ and $x_0 \in X$ such that $\|x_* - x_0\| \le R$, along with problem parameters $\varkappa, \varkappa', \nu, \varsigma_*^2, \underline{\kappa}$ and an upper bound $\bar{s}$ for signal sparsity. We are using the Stochastic Mirror Descent algorithm and apply the multistage modification of [27, 30] to improve its accuracy bounds. The proposed Stochastic Mirror Descent algorithm for Sparse Recovery (SMD-SR) works in stages—runs of the Stochastic Mirror Descent algorithm followed by subsequent "sparsification" of the approximate solution delivered by the SMD. The stages are split into two groups—phases—corresponding to two different regimes of the method. This organization of the algorithm allows to treat differently two components in the bound (14) for the error of the Stochastic Mirror Descent algorithm.

During the first *preliminary* phase of the algorithm, the first term in the right-hand side of (14) is dominant. This term is proportional to the bound R^2 on the expected squared ℓ_1-norm of the error of the initial solution, and decreases as $1/m$ where m is the iteration count. During the stages of the preliminary phase, the stepsize parameter β and the number of iterations per stage are set constant in such a way that the bound for the expected squared error of the approximate solution decreases by a constant factor at the end of the stage. Therefore, during this phase, the error of approximate solution converges linearly as a function of the total number of calls to stochastic oracle.

Preliminary phase terminates when the first term in the error bound (14) becomes dominated with the second, independent of the initial error of the algorithm. During the second *asymptotic* phase of the method, the choice of the stepsize parameter and the length of the stage are "standard" for multistage Stochastic Mirror Descent (cf., e.g., [30]) and the method converges sublinearly, with the "standard" rate $O(1/N)$ where N is the total number of oracle calls.

Algorithm 1 [SMD-SR]

1. Preliminary phase

   Initialization: Set $y_0 = x_0 \in X$, $R_0 = R$,

   $$\beta_0 = 2\varkappa\nu, \quad m_0 = \lceil 16\underline{\kappa}^{-1}\bar{s}(8\Theta\varkappa + 1)\nu \rceil \tag{15}$$

   (here $\lceil a \rceil$ stands for the smallest integer greater or equal to a). Put

   $$\overline{K} = \left\lceil \ln_2\left(\frac{R_0^2\underline{\kappa}\nu\varkappa}{32\varsigma_*^2\bar{s}\varkappa'}\right)\right\rceil$$

   and run

   $$K = \min\left\{\left\lfloor\frac{N}{m_0}\right\rfloor, \overline{K}\right\}$$

stages of the preliminary phase (here $\lfloor a \rfloor$ stands for the "usual" integer part – the largest integer less or equal to a).

Stage $k = 1, ..., K$*:* Compute approximate solution $\widehat{x}_{m_0}(y_{k-1}, \beta_0)$ after m_0 iterations of SMD algorithm with constant stepsize parameter β_0, corresponding to the initial condition $x_0 = y_{k-1}$. Then define y_k as "s-sparsification" of $\widehat{x}_{m_0}(y_{k-1}, \beta_0)$, i.e., $y_k = \mathrm{sparse}(\widehat{x}_{m_0}(y_{k-1}, \beta_0))$.

Output: define $\widehat{y}^{(1)} = y_K$ and $\widehat{x}^{(1)} = \widehat{x}_{m_0}(y_{K-1}, \beta)$ as approximate solutions at the end of the phase.

2. Set $M = N - m_0\overline{K}$ and

$$m_k = \left\lceil 512 \frac{\bar{s}\Theta\nu\varkappa}{\underline{\kappa}} 2^k \right\rceil, \quad k = 1, ...$$

If $m_1 > M$ terminate and output $\widehat{y}_N = \widehat{y}^{(1)}$ and $\widehat{x}_N = \widehat{x}^{(1)}$ as approximate solutions by the procedure; otherwise, continue with stages of the asymptotic phase.

Asymptotic phase

Initialization: Set

$$K' = \max\left\{k : \sum_{i=1}^{k} m_i \le M\right\},$$

$y'_0 = \widehat{y}^{(1)}$, and $\beta_k = 2^k\nu\varkappa$, $k = 1, ..., K'$.

Stage $k = 1, ..., K'$*:* Compute $\widehat{x}_{m_k}(y'_{k-1}, \beta_k)$; same as above, define $y'_k = \mathrm{sparse}(\widehat{x}_{m_k}(y'_{k-1}, \beta_k))$.

Output: After K' stages, output $\widehat{y}_N = y'_{K'}$ and $\widehat{x}_N = \widehat{x}_{m_{K'}}(y'_{K'-1}, \beta_{K'})$.

Properties of the proposed procedure are summarized in the following statement.

Theorem 2.1 *In the situation of this section, suppose that $N \ge m_0$ so at least one preliminary stage of Algorithm 1 is completed. Then approximate solutions $\widehat{x}_N$ and $\widehat{y}_N$ produced by the algorithm satisfy*

$$\begin{aligned}
\mathrm{Risk}_g(\widehat{x}_N|\mathcal{X}) &\le \frac{\kappa R^2}{\bar{s}} \exp\left\{-\frac{cN\underline{\kappa}}{\Theta\varkappa\bar{s}\nu}\right\} + C\frac{\varsigma_*^2\bar{s}\varkappa'\Theta}{\underline{\kappa}N}, &(16)\\
\mathrm{Risk}_{\|\cdot\|}(\widehat{y}_N|\mathcal{X}) &\le \sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2}(\widehat{y}_N|\mathcal{X}) \le \sqrt{8s}\mathrm{Risk}_{\|\cdot\|_2}(\widehat{x}_N|\mathcal{X})\\
&\lesssim R\exp\left\{-\frac{cN\underline{\kappa}}{\Theta\varkappa\bar{s}\nu}\right\} + \frac{\varsigma_*\bar{s}}{\underline{\kappa}}\sqrt{\frac{\Theta\varkappa'}{N}}. &(17)
\end{aligned}$$

2.4 Enhancing the reliability of SMD-SR solutions

In this section, our objective is to build approximate solutions to problem (5) utilizing Algorithm 1 which obey "sub-Gaussian type" bounds on their ϵ-risks. Note that bounds (16) and (17) of Theorem 2.1 do allow only for Chebyshev-type bounds for risks of $\widehat{y}_N$ and $\widehat{x}_N$. Nevertheless, their confidence can be easily improved by applying, for instance, an adapted version of "median-of-means" estimate [42, 46].

Reliable recovery utilizing geometric median of SMD-SR solutions. Suppose that available sample of length N can be split into L independent samples of length $M = N/L$ (for the sake of simplicity let us assume that N is a multiple of L). We run Algorithm 1 on each subsample thus obtaining L independent recoveries $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$ and compute "enhanced solutions" using an aggregation procedure of geometric median-type. Note that we are in the situation where Theorem 2.1 applies, meaning that approximate solutions $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$ satisfy

$$\forall \ell \quad \mathbf{E}\{g(\widehat{x}_M^{(\ell)})\} - g_* \le \tau_M^2 := \frac{\underline{\kappa} R^2}{\bar{s}} \exp\left\{-\frac{cM\underline{\kappa}}{\Theta \varkappa \bar{s} \nu}\right\} + C\frac{\varsigma_*^2 \bar{s} \varkappa' \Theta}{\underline{\kappa} M}, \tag{18}$$

and so

$$\forall \ell \quad \mathbf{E}\{\|\widehat{x}_M^{(\ell)} - x_*\|_2^2\} \le \theta_M^2 := \frac{2}{\underline{\kappa}} \tau_M^2 \lesssim \frac{R^2}{\bar{s}} \exp\left\{-\frac{cM\underline{\kappa}}{\Theta \varkappa \bar{s} \nu}\right\} + \frac{\Theta \varkappa' \varsigma_*^2 \bar{s}}{\underline{\kappa}^2 M}. \tag{19}$$

We are to select among $\widehat{x}_M^{(\ell)}$ the solution which attains similar bounds "reliably."

1. The first reliable solution $\widehat{x}_{N,1-\epsilon}$ of x_* is a "pure" geometric median of $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$: we put

$$\widehat{x}_{N,1-\epsilon} \in \operatorname*{Argmin}_x \sum_{\ell=1}^{L} \|x - \widehat{x}_M^{(\ell)}\|_2, \tag{20}$$

   and then define $\widehat{y}_{N,1-\epsilon} = \mathrm{sparse}(\widehat{x}_{N,1-\epsilon})$. [7]

   Computing reliable solutions $\widehat{x}_{N,1-\epsilon}$ and $\widehat{y}_{N,1-\epsilon}$ as optimal solutions to (20) amounts to solving a nontrivial optimization problem. A simpler reliable estimation can be computed by replacing the geometric median $\widehat{x}_{N,1-\epsilon}$ by its "empirical counterparts" (note that, number L of solutions to be aggregated is not large—it is typically order of $\ln[1/\epsilon]$).

2. We can replace $\widehat{x}_{N,1-\epsilon}$ with

$$\widehat{x}'_{N,1-\epsilon} \in \operatorname*{Argmin}_{x \in \{\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}\}} \sum_{\ell=1}^{L} \|x - \widehat{x}_M^{(\ell)}\|_2$$

   and compute its sparse approximation $\widehat{y}'_{N,1-\epsilon} = \mathrm{sparse}(\widehat{x}'_{N,1-\epsilon})$.

3. Another reliable solution (with slightly better guarantees) was proposed in [25]. Let $i \in \{1, ..., L\}$, we set

$$r_{ij} = \|\widehat{x}_M^{(i)} - \widehat{x}_M^{(j)}\|_2$$

   and denote $r_{(1)}^i \le r_{(2)}^i \le ... \le r_{(L-1)}^i$ corresponding order statistics (i.e., $r_{i\cdot}$'s sorted in the increasing order). We define reliable solution $\widehat{x}''_{N,1-\epsilon} = \widehat{x}_M^{(\widehat{i})}$ where

$$\widehat{i} \in \operatorname*{Argmin}_{i \in \{1,...,L\}} r^i_{\lceil L/2 \rceil} \tag{21}$$

   (here $\lceil a \rceil = \lfloor a \rfloor + 1$ stands for the smallest integer strictly greater than a), and put $\widehat{y}''_{N,1-\epsilon} = \mathrm{sparse}(\widehat{x}''_{N,1-\epsilon})$.

[7] Reliable solution we consider here explicitly depend on the confidence level; for instance, parameter L in the definition (20) of $\widehat{x}_{N,1-\epsilon}$ will be chosen depending on ϵ. Hence, the presence of the index $1-\epsilon$ in the notation of these estimates.

Theorem 2.2 *Let $\epsilon \in (0, \frac{1}{4}]$, and let $\overline{x}_N$ (resp. $\overline{y}_N$) be one of reliable solutions $\widehat{x}_{N,1-\epsilon}, \widehat{x}'_{N,1-\epsilon}$ and $\widehat{x}''_{N,1-\epsilon}$ (resp., $\widehat{y}_{N,1-\epsilon}, \widehat{y}'_{N,1-\epsilon}$ and $\widehat{y}''_{N,1-\epsilon}$) described above using $L = \lceil \alpha \ln[1/\epsilon] \rceil$[8] independent approximate solutions $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$ by Algorithm 1. When $N \geq Lm_0$ we have*

$$\begin{aligned} \mathrm{Risk}_{\|\cdot\|,\epsilon}(\overline{y}_N|\mathcal{X}) &\leq \sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2,\epsilon}(\overline{y}_N|\mathcal{X}) \leq 2\sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2,\epsilon}(\overline{x}_N|\mathcal{X}) \\ &\lesssim R\exp\left\{-\frac{cN\underline{\kappa}}{\Theta\varkappa\bar{s}\nu\ln[1/\epsilon]}\right\} + \frac{\varsigma_*\bar{s}}{\underline{\kappa}}\sqrt{\frac{\Theta\varkappa'\ln[1/\epsilon]}{N}}. \end{aligned} \tag{22}$$

Remark. Notice that the term $\ln[1/\epsilon]$ enters the bound (22) as a multiplier which is typical for accuracy estimates of solutions which relies upon median to enhance confidence; at the moment, we do not know if this dependence on reliability tolerance parameter may be improved.

Reliable solution aggregation. Let us assume that two independent observation samples of lengths N and K are available. In the present approach, we use the first sample to compute, same as in the construction presented above, L independent approximate SMD-SR solutions $\widehat{x}_M^{(\ell)}$, $\ell = 1, ..., L$, $M = N/L$. Then we "aggregate" $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$—select the best of them in terms of the objective value $g(\widehat{x}_M^{(\ell)})$ by computing reliable estimations of differences $g(\widehat{x}_M^{(i)}) - g(\widehat{x}_M^{(j)})$ using observations of the second subsample.

The proposed procedure for reliable selection of the "best" solution $\widehat{x}_M^{(\ell)}$ is as follows.

Algorithm 2 [Reliable aggregation]

Initialization: Algorithm parameters are $\epsilon \in (0, \frac{1}{2}]$, $L' \in \mathbf{Z}_+$ and $m = K/L'$ (for the sake of simplicity we assume, as usual, that $K = mL'$). We assume to be given L points $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$ (approximate solution of the first step).
We compute $\widehat{x}''_{N,1-\epsilon} = \widehat{x}_M^{(\widehat{i})}$ the reliable solution as defined in (21) and denote $\widehat{I} = \{i_1, ..., i_{\lceil L/2 \rceil}\}$, the set of indices of $\lceil L/2 \rceil$ closest to $\widehat{x}''_{N,1-\epsilon}$ in the Euclidean norm points among $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$.

Comparison procedure: We split the (second) sample ω^K into L' independent subsamples ω^ℓ, $\ell = 1, ..., L'$ of size m. For all $i \in \widehat{I}$ we compute the index

$$\widehat{v}_i = \max_{j \in \widehat{I}, j \neq i} \left\{ \operatorname*{median}_\ell [\widehat{v}_{ji}^\ell] - \rho_{ij} \right\}$$

where

$$\widehat{v}_{ji}^\ell = \frac{1}{m}\sum_{k=1}^m \left\langle \nabla G(\widehat{x}_M^{(j)} + t_k(\widehat{x}_M^{(i)} - \widehat{x}_M^{(j)}), \omega_k^\ell), \widehat{x}_M^{(i)} - \widehat{x}_M^{(j)} \right\rangle, \quad \ell = 1, ..., L',$$

are estimates of $v_{ji} = g(\widehat{x}_M^{(i)}) - g(\widehat{x}_M^{(j)})$, $t_k = \frac{2k-1}{2m}$, $k = 1, ..., m$, and coefficients $\rho_{ij} > 0$ to be defined depend on $r_{ij} = \|\widehat{x}_M^{(i)} - \widehat{x}_M^{(j)}\|_2$.

- Output: We say that $x_M^{(i)}$ *is admissible* if $\widehat{v}_i \leq 0$. When the set of admissible $\widehat{x}_M^{(i)}$'s is nonempty we define the procedure output $\overline{x}_{N+K,1-\epsilon}$ as one of admissible $\widehat{x}_M^{(i)}$'s, and define $\overline{x}_{N+K,1-\epsilon} = \widehat{x}_M^{(1)}$ otherwise.

Now, consider the following (cf. Assumption **S1**)

[8] The exact value of the numeric constant α is specific for each construction, and can be retrieved from the proof of the theorem.

Assumption [S3]. *There are $1 \le \chi, \chi' < \infty$ such that for any $x \in X$ and $z \in E$ the following bound holds:*

$$\mathbf{E}\{\langle \zeta(x,\omega), z\rangle^2\} \le \|z\|_2^2[\chi \mathcal{L}_2(g(x) - g_*) + \chi' \varsigma_*^2] \tag{23}$$

where $\mathcal{L}_2$ is the Lipschitz constant of the gradient ∇g of g with respect to the Euclidean norm,

$$\|\nabla g(x') - \nabla g(x'')\|_2 \le \mathcal{L}_2 \|x' - x''\|_2, \quad \forall x', x'' \in X.$$

Let now $\overline{\mathcal{X}}$ be the class of Sparse Stochastic Optimization problems as described in Section 2.1 satisfying Assumptions **S1**–**S3**, with domain X. Assume that risk $\mathrm{Risk}_{g,\epsilon}(\cdot|\overline{\mathcal{X}})$ is defined as in (10) with $\mathcal{X}$ replaced with $\overline{\mathcal{X}}$.

Theorem 2.3 *Let Assumption* **S3** *hold, and let τ_M and θ_M be as in (18) and (19) respectively. Further, in the situation of this section, let $\epsilon \in (0, \frac{1}{2}]$, $L = \lceil \alpha \ln[1/\epsilon] \rceil$ for large enough α, and let $\overline{x}_{N+K,1-\epsilon}$ be an approximate solution by Algorithm 2 in which we set $L' \ge \left\lceil 7 \ln[2/\varepsilon] \right\rceil$ and*

$$\rho_{ij} = 2r_{ij}\sqrt{\frac{\mathcal{L}_2\chi}{m}}(\gamma(r_{ij}) + \tau_M) + 2r_{ij}\varsigma_*\sqrt{\frac{\chi'}{m}}$$

where

$$\gamma(r) = \left(\left[4r\sqrt{\frac{\chi\mathcal{L}_2}{m}} + \tau_M\right]^2 + 4r\zeta_*\sqrt{\frac{\chi'}{m}}\right)^{1/2}. \tag{24}$$

Then

$$\mathrm{Risk}_{g,\epsilon}(\overline{x}_{N+K,1-\epsilon}|\overline{\mathcal{X}}) \le \bar{\gamma}^2 := \gamma^2(8\theta_M),$$

In particular, when $K = mL' \ge c\max\left\{\frac{\chi\mathcal{L}_2 \ln[1/\epsilon]}{\underline{\kappa}}, \frac{N\chi'}{\Theta\varkappa'\bar{s}}\right\}$ for an appropriate absolute $c > 0$, one has

$$\mathrm{Risk}_{g,\epsilon}(\overline{x}_{N+K,1-\epsilon}|\overline{\mathcal{X}}) \lesssim \frac{\underline{\kappa}R^2}{\bar{s}}\exp\left\{-\frac{cN\underline{\kappa}}{\Theta\varkappa\bar{s}\nu\ln[1/\epsilon]}\right\} + \frac{\varsigma_*^2\bar{s}\Theta\varkappa'\ln[1/\epsilon]}{\underline{\kappa}N}.$$

3 Applications

3.1 Sparse generalized linear regression by stochastic approximation

Let us consider the problem of recovery of a sparse signal $x_* \in \mathbf{R}^n$, $n \ge 3$, from independent and identically distributed observations

$$\eta_i = \mathfrak{u}(\phi_i^T x_*) + \sigma\xi_i, \quad i = 1, 2, ..., N, \tag{25}$$

where "activation" $u : \mathbf{R} \to \mathbf{R}$, $\phi_i \in \mathbf{R}^n$ and $\xi_i \in \mathbf{R}$ are mutually independent and such that $\mathbf{E}\{\phi_i\phi_i^T\} = \Sigma$, $\kappa_\Sigma I \preceq \Sigma$, and $\|\Sigma\|_\infty \le \upsilon$, with known $\kappa_\Sigma > 0$ and υ;[9] we also assume that $\mathbf{E}\{\xi_i\} = 0$ and $\mathbf{E}\{\xi_i^2\} \le 1$.

We suppose that x_* is *s-sparse* and that we are given a convex and closed subset X of $\mathbf{R}^n$ (e.g., a large enough ball of ℓ_1- or ℓ_2-norm centered at the origin) such that $x_* \in X$, along with $R < \infty$ and $x_0 \in X$ such that $\|x_* - x_0\|_1 \le R$. Furthermore, the mapping $\mathfrak{u}(\cdot)$ is assumed to be known, strongly monotone and Lipschitz continuous, i.e., for some $0 < \underline{\ell} \le \bar{\ell}$ and all $t \ge t'$

$$\underline{\ell}(t - t') \le \mathfrak{u}(t) - \mathfrak{u}(t') \le \bar{\ell}(t - t'). \tag{26}$$

[9] Recall that for a matrix Q we denote $\|Q\|_\infty = \max_{ij} |[Q]_{ij}|$.

We are about to apply Stochastic Optimization approach described in Section 2. To this end, let $\mathfrak{v}$ be the primitive of u, i.e., $\mathfrak{v}'(t) = \mathfrak{u}(t)$, and let us consider the Stochastic Optimization problem

$$\min_{x\in X}\left\{g(x) = \tfrac{1}{2}\mathbf{E}\big\{\underbrace{\mathfrak{v}(\phi^T x) - \phi^T x\eta}_{=:G(x,\omega=[\phi,\eta])}\big\}\right\}. \tag{27}$$

Note that x_* is the unique optimal solution to the above problem. Indeed, observe that $\nabla G(x,\omega) = \phi(\mathfrak{u}(\phi^T x) - \eta)$ and $\mathbf{E}_\xi\{\eta\} = \mathfrak{u}(\phi^T x_*)$. We have $\nabla g(x_*) = 0$; furthermore,

$$\begin{array}{rcl} g(x) - g(x_*) & = & \displaystyle\int_0^1 \nabla g(x_* + t(x-x_*))^T(x-x_*)dt \\ & = & \displaystyle\int_0^1 \mathbf{E}\Big\{\phi[\mathfrak{u}(\phi^T(x_*+t(x-x_*)) - \mathfrak{u}(\phi^T x_*)]\Big\}^T(x-x_*)dt \\ [\text{by (26)}] & \geq & \displaystyle\int_0^1 \underline{\ell}\mathbf{E}\big\{[\phi^T(x-x_*)]^2\big\}tdt = \tfrac{1}{2}\underline{\ell}\|x-x_*\|_\Sigma^2 \geq \tfrac{1}{2}\underline{\ell}\kappa_\Sigma\|x-x_*\|_2^2, \end{array}$$

and we conclude that g is quadratically minorated with parameter $\underline{\kappa} = \underline{\ell}\kappa_\Sigma$.

We set $\|\cdot\| = \|\cdot\|_1$ with $\|\cdot\|_* = \|\cdot\|_\infty$, and we use "$\ell_1$-proximal setup" of the SMD-SR algorithm with quadratically growing for $n > 2$ distance-generating function (cf. [49, Theorem 2.1])

$$\vartheta(x) = \tfrac{1}{2}e\ln(n)\, n^{(p-1)(2-p)/p}\|x\|_p^2, \quad p = 1 + \frac{1}{\ln n},$$

the corresponding Θ satisfying $\Theta \leq \frac{1}{2}e^2\ln n$.

Note that, due to (26), for all $z \in \mathbf{R}^n$ such that $\|z\|_1 \leq 1$

$$\begin{array}{rcl} |z^T(\nabla g(x) - \nabla g(x'))| & = & \big|\mathbf{E}\big\{\phi^T z(\mathfrak{u}(\phi^T x) - \mathfrak{u}(\phi^T x'))\big\}\big| \leq \overline{\ell}\mathbf{E}\{|\phi^T z|\,|\phi^T(x-x')|\} \\ & \leq & \overline{\ell}\mathbf{E}\{(\phi^T z)^2\}^{1/2}\mathbf{E}\big\{(\phi^T(x-x'))^2\big\}^{1/2} \leq \overline{\ell}\upsilon^{1/2}\|x-x'\|_\Sigma, \end{array}$$

i.e., $\|\nabla g(x) - \nabla g(x')\|_\infty \leq \overline{\ell}\upsilon^{1/2}\|x-x'\|_\Sigma$. Thus,

$$\begin{array}{rcl} \varsigma(x) & = & \mathbf{E}\big\{\|\nabla G(x,\omega) - \nabla g(x)\|_\infty^2\big\}^{1/2} \leq \mathbf{E}\big\{[\|\phi(\mathfrak{u}(\phi^T x) - \mathfrak{u}(\phi^T x_*)) - \nabla g(x)\|_\infty + \|\phi\xi\|_\infty]^2\big\}^{1/2} \\ & \leq & \overline{\ell}\mathbf{E}\big\{\|\phi\|_\infty^2(\phi^T(x-x_*))^2\big\}^{1/2} + \overline{\ell}\upsilon^{1/2}\|x-x'\|_\Sigma + \nu\sigma \end{array}$$

where $\nu = \mathbf{E}\{\|\phi\|_\infty^2\}^{1/2}$. In other words, Assumption **S1** holds whenever

$$\varsigma^2(x) \leq \Big(\overline{\ell}\mathbf{E}\big\{\|\phi\|_\infty^2(\phi^T(x-x_*))^2\big\}^{1/2} + \overline{\ell}\upsilon^{1/2}\|x-x'\|_\Sigma + \nu\sigma\Big)^2 \leq \varkappa\nu(g(x) - g_*) + \varkappa'\varsigma_*^2 \tag{28}$$

which is the case if, for instance,

$$\overline{\ell}^2\mathbf{E}\big\{\|\phi\|_\infty^2(\phi^T(x-x_*))^2\big\} \lesssim \nu\underline{\ell}\|x-x'\|_\Sigma^2. \tag{29}$$

and ς_* satisfies $\varsigma_*^2 \geq \nu^2\sigma^2$.

Remark. In the special case of $\mathfrak{u}(t) = t$, one has

$$\begin{aligned} g(x) &= \mathbf{E}\big\{\underbrace{\tfrac{1}{2}(\phi^T x)^2 - \phi^T x\eta}_{=G(x,\omega)}\big\} = \tfrac{1}{2}\mathbf{E}\big\{[\phi^T(x_* - x)]^2 - (\phi^T x_*)^2\big\} \\ &= \tfrac{1}{2}(x - x_*)^T\Sigma(x - x_*) - \tfrac{1}{2}x_*^T\Sigma x_* = \tfrac{1}{2}\|x - x_*\|_\Sigma^2 - \tfrac{1}{2}\|x_*\|_\Sigma^2 \end{aligned}$$

with $\nabla g(x) = \Sigma(x - x_*) = \mathbf{E}\big\{\underbrace{\phi\phi^T(x - x_*) - \sigma\xi\phi}_{=:\nabla G(x,\omega)}\big\}$. In this case,

$$\zeta(x,\omega) = \nabla G(x,\omega) - \nabla g(x) = [\phi\phi^T - \Sigma](x - x_*) - \sigma\xi\phi,$$

and

$$\varsigma^2(x) = \mathbf{E}\big\{\|[\phi\phi^T - \Sigma](x - x_*) - \sigma\xi\phi\|_\infty^2\big\}.$$

In this situation, Assumption **S1** simplifies to

$$\mathbf{E}\big\{\|[\phi\phi^T - \Sigma](x - x_*) - \sigma\xi\phi\|_\infty^2\big\} \le \tfrac{1}{2}\varkappa\nu\|x - x_*\|_\Sigma^2 + \varkappa'\varsigma_*^2$$

which is satisfied with $\varsigma_*^2 = \nu^2\sigma^2$ whenever $\mathbf{E}\big\{\|\phi\|_\infty^2(\phi^T(x - x_*))^2\big\} \lesssim \nu\|x - x'\|_\Sigma^2$.

Our present goal is to describe the properties of approximate solutions by Algorithm 1 when applied to the optimization problem in (27). We assume that the problem parameters—values $\varkappa$, ν, κ_Σ, σ^2 and an upper bound $\bar{s}$ on sparsity of x_*—are known. We consider the following performance characteristics of approximate solutions $\widehat{x}$—analogues of risks measures defined in Section 2.1—in our present situation:

- *Recovery risks:* maximal over $x_* \in X$ expected squared error

$$\mathrm{Risk}_{|\cdot|}(\widehat{x}|X) = \sup_{x_*\in X} \mathbf{E}\{|\widehat{x} - x_*|^2\}^{1/2} \tag{30}$$

  where $|\cdot|$ stands for $\|\cdot\|_2$- or $\|\cdot\|$-norm (which is $\|\cdot\|_1$-norm in the sparse regression setting), and ϵ-*risk* of recovery—the smallest maximal over $x_* \in X$ radius of $(1-\epsilon)$-confidence ball of norm $|\cdot|$ centered at $\widehat{x}$:

$$\mathrm{Risk}_{|\cdot|,\epsilon}(\widehat{x}|X) = \inf\left\{r : \sup_{x_*\in X} \mathrm{Prob}\{|\widehat{x} - x_*| \ge r\} \le \epsilon\right\} \tag{31}$$

- *Prediction risks:* maximal over $x_* \in X$ expected suboptimality

$$\mathrm{Risk}_g(\widehat{x}|X) = \sup_{x_*\in X} \mathbf{E}\{g(\widehat{x})\} - g_*, \tag{32}$$

  of $\widehat{x}$ and the smallest maximal over $x_* \in X$ $(1-\epsilon)$-confidence interval

$$\mathrm{Risk}_{g,\epsilon}(\widehat{x}|X) = \inf\left\{r : \sup_{x_*\in X} \mathrm{Prob}\{g(\widehat{x}) - g_* \ge r\} \le \epsilon\right\}. \tag{33}$$

The following statement is a straightforward corollary of Theorems 2.1 and 2.2.

Proposition 3.1 *Suppose that (28) holds.*

(i) Let the sample size N satisfy

$$N \geq m_0 = \left\lceil \frac{16\nu\bar{s}}{\underline{\ell}\kappa_\Sigma}(4e^2\varkappa\ln[n]+1)\right\rceil$$

so at least one preliminary stage of Algorithm 1 is completed. Then approximate solutions $\widehat{x}_N$ and $\widehat{y}_N$ produced by the algorithm satisfy

$$\begin{aligned}\mathrm{Risk}_{\|\cdot\|}(\widehat{y}_N|X) &\leq 2\sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2}(\widehat{x}_N|X) \lesssim R\exp\left\{-\frac{cN\underline{\ell}\kappa_\Sigma}{\varkappa\bar{s}\nu\ln n}\right\}+\frac{\sigma\bar{s}}{\underline{\ell}\kappa_\Sigma}\sqrt{\frac{\nu\ln n}{N}} \qquad (34)\\ \mathrm{Risk}_g(\widehat{x}_N|X) &\lesssim \frac{\underline{\ell}\kappa_\Sigma R^2}{\bar{s}}\exp\left\{-\frac{cN\underline{\ell}\kappa_\Sigma}{\varkappa\bar{s}\nu\ln n}\right\}+\frac{\nu\sigma^2\bar{s}\varkappa'\ln n}{\underline{\ell}\kappa_\Sigma N}.\end{aligned}$$

(ii) Furthermore, when observation size satisfies $N \geq \alpha m_0\ln[1/\epsilon]$ with large enough absolute $\alpha > 0$, $1-\epsilon$ reliable solutions $\widehat{y}_{N,1-\epsilon}$ and $\widehat{x}_{N,1-\epsilon}$ as defined in Section 2.4 satisfy

$$\begin{aligned}\mathrm{Risk}_{\|\cdot\|,\epsilon}(\widehat{y}_{N,1-\epsilon}|X) &\leq \sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2,\epsilon}(\widehat{y}_{N,1-\epsilon}|X) \leq 2\sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2,\epsilon}(\widehat{x}_{N,1-\epsilon}|X)\\ &\lesssim R\exp\left\{-\frac{cN\underline{\ell}\kappa_\Sigma}{\varkappa\bar{s}\nu\ln[1/\epsilon]\ln n}\right\}+\frac{\sigma\bar{s}}{\underline{\ell}\kappa_\Sigma}\sqrt{\frac{\nu\ln[1/\epsilon]\ln n}{N}}, \qquad (35)\end{aligned}$$

with $\widehat{x}'_{N,1-\epsilon}$, $\widehat{x}''_{N,1-\epsilon}$ and $\widehat{y}'_{N,1-\epsilon}$, $\widehat{y}''_{N,1-\epsilon}$ verifying similar bounds.

Let $\sigma_1(\Sigma)$ be the principal eigenvalue (the spectral norm) of Σ. Then, for all z such that $\|z\|_2 = 1$ one has

$$\begin{aligned}z^T(\nabla g(x)-\nabla g(x')) &= \mathbf{E}\{z^T\phi(\mathfrak{u}(\phi^Tx)-\mathfrak{u}(\phi^Tx'))\} \leq \bar{\ell}\mathbf{E}\{|z^T\phi|\,|\phi^T(x-x')|\}\\ &\leq \bar{\ell}\mathbf{E}\{(\phi^Tz)^2\}^{1/2}\mathbf{E}\{(\phi^T(x-x'))^2\}^{1/2} \leq \bar{\ell}\sigma_1(\Sigma)\|x-x'\|_2,\end{aligned}$$

implying that the Lipschitz constant of ∇g with respect to the Euclidean norm can be set as $\mathcal{L}_2 = \bar{\ell}\sigma_1(\Sigma)$. Thus, Assumption **S3** holds when for some $1 \leq \chi < \infty$ and all $x \in X$, $z \in \mathbf{R}^n$

$$\underline{\ell}^{-1}\bar{\ell}\,\mathbf{E}\left\{\left(z^T\phi\right)^2\left(\phi^T(x-x_*)\right)^2\right\} \leq \tfrac{1}{2}\chi\|z\|_2^2\sigma_1(\Sigma)\|x-x_*\|_\Sigma^2. \qquad (36)$$

Indeed, in this case one has for all $z \in \mathbf{R}^n$:

$$\begin{aligned}\mathbf{E}\left\{(z^T\zeta(x,\omega))^2\right\} &= \mathbf{E}\left\{(z^T\phi)^2\left[(\mathfrak{u}(\phi^Tx)-\mathbf{E}\{\mathfrak{u}(\phi^Tx)\})-(\mathfrak{u}(\phi^Tx_*)-\mathbf{E}\{\mathfrak{u}(\phi^Tx_*)\})-\sigma\xi\right]^2\right\}\\ &\leq \mathbf{E}\left\{\left(z^T\phi\right)^2\left(\mathfrak{u}(\phi^Tx)-\mathfrak{u}(\phi^Tx_*)\right)^2\right\}+\sigma^2\mathbf{E}\{\xi^2(\phi^Tz)^2\}\\ &\leq \bar{\ell}^2\mathbf{E}\left\{\left(z^T\phi\right)^2\left(\phi^T(x-x_*)\right)^2\right\}+\sigma^2\|z\|_2^2\sigma_1(\Sigma)\\ [\text{by (36)}] &\leq \tfrac{1}{2}\bar{\ell}\underline{\ell}\|x-x_*\|_\Sigma^2\chi\|z\|_2^2\sigma_1(\Sigma)+\sigma^2\|z\|_2^2\sigma_1(\Sigma)\\ &\leq (g(x)-g_*)\chi\|z\|_2^2\mathcal{L}_2+\sigma_1(\Sigma)\sigma^2\|z\|_2^2\end{aligned}$$

implying (23) with $\chi' = \sigma_1(\Sigma)/\nu^2$.

The following result is a corollary of Theorem 2.3.

Proposition 3.2 *Suppose that (28) and (36) hold true, and let*

$$N \geq c\max\left\{\frac{\varkappa\nu\bar{s}}{\underline{\ell}\kappa_\Sigma}\ln[1/\epsilon]\ln n,\ \frac{\chi\sigma_1(\Sigma)}{\underline{\ell}\kappa_\Sigma}\ln[1/\epsilon]\right\}$$

with large enough $c > 0$. Then aggregated solution $\overline{x}_{2N,1-\epsilon}$ (with $K = N$) by Algorithm 2 satisfies

$$\mathrm{Risk}_{g,\epsilon}(\overline{x}_{2N,1-\epsilon}|X) \lesssim \frac{\underline{\ell}\kappa_\Sigma R^2}{\bar{s}}\exp\left\{-\frac{cN\underline{\ell}\kappa_\Sigma}{\varkappa\bar{s}\nu\ln[1/\epsilon]\ln n}\right\}+\frac{\sigma^2\nu\bar{s}\ln[1/\epsilon]\ln n}{\underline{\ell}\kappa_\Sigma N}. \qquad (37)$$

Note that when $\sigma_1(\Sigma) = O(\nu\ln n)$ and $\varkappa$ and χ are both $O(1)$ bounds (35) and (37) hold for $N \geq c\frac{\nu\bar{s}}{\underline{\ell}\kappa_\Sigma}\ln[1/\epsilon]\ln n$.

Remark. Results of Propositions 3.1 and 3.2 merit some comments. If compared to now standard accuracy bounds for sparse recovery by ℓ_1-minimization [3, 9, 10, 13, 28, 52, 54, 57], to the best of our knowledge, (28) and (36) provide the most relaxed conditions under which the bounds such as (34)–(37) can be established. An attentive reader will notice a degradation of bounds (35) and (37) with respect to comparable results [19, 28, 52] as far as dependence in factors which are logarithmic in n and ϵ^{-1} is concerned—bound (22) depends on the product $\ln[n]\ln[1/\epsilon]$ of these terms instead of the sum $\ln[n]+\ln[\epsilon^{-1}]$ in the "classical" results.[10] This seems to be a technical "artifact" of the analysis of non-Euclidean stochastic approximation algorithm and the reliability enhancement approach using median of estimators we have adopted in this work, cf. the comment after Theorem 2.2. Nevertheless, it is rather surprising to see that conditions on the regressor model in Proposition 3.1, apart from positive definiteness of regressor covariance matrix, essentially amount to (cf. (29))

$$\mathbf{E}\big\{\|\phi\|_\infty^2(\phi^Tz)^2\big\}\lesssim\nu\|z\|_\Sigma^2\ \ \forall z\in\mathbf{R}^n.$$

Below we consider some examples of situations where bounds (29) and (36) hold with constants which are "almost" dimension-independent, i.e. are, at most, *logarithmic in problem dimension.* When this is the case, and when observation count N satisfies $N\geq\alpha m_0\ln[1/\epsilon]\ln[R/(s\sigma)]$ for large enough absolute α, so that the preliminary phase of the algorithm is completed, the bounds of Propositions 3.1 and 3.2 coincide (up to already mentioned logarithmic in n and $1/\epsilon$ factors) with the best accuracy bound available for sparse recovery in the situation in question.[11]

1. *Sub-Gaussian regressors:* suppose now that $\phi_i\sim\mathrm{Sub}\mathcal{G}(0,S)$, i.e., regressors ϕ_i are sub-Gaussian with zero mean and matrix parameter S, meaning that

$$\mathbf{E}\big\{e^{u^T\phi}\big\}\leq e^{\frac{u^TSu}{2}}\ \text{ for all } u\in\mathbf{R}^n.$$

Let us assume that sub-Gaussianity matrix S is "similar" to the covariance matrix Σ of ϕ, i.e. $S\preceq\mu\Sigma$ with some $\mu<\infty$. Note that $\mathbf{E}\{(\phi^Tz)^4\}\leq16(z^TSz)^2\leq16\mu^2\|z\|_\Sigma^4$, and thus

$$\mathbf{E}\{(z^T\phi\phi^Tx)^2\}\leq\mathbf{E}\{(z^T\phi)^4\}^{1/2}\mathbf{E}\{(x^T\phi)^4\}^{1/2}\leq16z^TSz\,x^TSx\leq16\mu^2\sigma_1(\Sigma)\|z\|_2^2\|x\|_\Sigma^2,$$

which is (36) with $\chi=16\mu^2\underline{\ell}^{-1}\overline{\ell}$. Let us put $\bar{v}=\max_i[S]_{ii}$. One easily verifies that in this case

$$\nu^2=\mathbf{E}\{\|\phi\|_\infty^2\}\leq2\bar{v}(\ln[2n]+1)\leq2\mu v(\ln[2n]+1),$$

and

$$\mathbf{E}\{\|\phi\|_\infty^4\}\leq4\bar{v}^2(\ln^2[2n]+2\ln[2n]+2)\leq4\mu^2v^2(\ln^2[2n]+2\ln[2n]+2).$$

As a result, we have, cf. (28),

$$\begin{aligned}\varsigma^2(x)&\leq\Big[\overline{\ell}(\mathbf{E}\{\|\phi\|_\infty^4\})^{1/4}(\mathbf{E}\{(\phi^T(x-x_*))^4\})^{1/4}+\sigma(\mathbf{E}\{\|\phi\|_\infty^2\})^{1/2}+\overline{\ell}\sqrt{v}\|x-x_*\|_\Sigma\Big]^2\\&\leq\Big[\overline{\ell}\sqrt{8\bar{v}(\ln[2n]+2)}\|x-x_*\|_S+\sigma\sqrt{2\bar{v}(\ln[2n]+1)}+\overline{\ell}\sqrt{v}\|x-x_*\|_\Sigma\Big]^2\\&\leq2\overline{\ell}^2\big(\mu\sqrt{8(\ln[2n]+2)}+1\big)^2v\|x-x_*\|_\Sigma^2+4\mu v(\ln[2n]+1)\sigma^2,\end{aligned}$$

whence, Assumption **S1** holds with $\varkappa\nu\lesssim\overline{\ell}^2\underline{\ell}^{-1}\mu^2v\ln n$, $\varkappa'\lesssim1$, and $\varsigma_*^2\lesssim\mu v\sigma^2\ln n$.

[10] Note that a similar deterioration was noticed in [13].

[11] In the case of "isotropic sub-Gaussian" regressors, see [38], the bounds of Proposition 3.1 are comparable to bounds of [37, Theorem 5] for Lasso recovery under relaxed moment assumptions on the noise ξ.

2. *Bounded regressors:* we assume that $\|\phi_i\|_\infty \le \mu$ a.s.. One has

$$\begin{aligned}\varsigma^2(x) &\le \left(\overline{\ell}\mu\mathbf{E}\{(\phi^T(x-x_*))^2\}^{1/2} + \overline{\ell}\upsilon^{1/2}\|x-x_*\|_\Sigma + \mu\sigma\right)^2 \\ &\le 2\overline{\ell}^2(\mu+\upsilon^{1/2})^2\|x-x_*\|_\Sigma^2 + 2\mu^2\sigma^2,\end{aligned}$$

implying the second inequality of (28) and also (8) with $\varkappa\nu \le 4\overline{\ell}^2\underline{\ell}^{-1}(\mu+\sqrt{\upsilon})^2$ and $\varsigma_*^2 \le \mu^2\sigma^2$. In particular, this condition is straightforwardly satisfied when ϕ_j are sampled from an orthogonal system with uniformly bounded elements, e.g., $\phi_j = \sqrt{n}\psi_{\kappa_j}$ where $\{\psi_j, j=1,...,n\}$ is a trigonometric or Hadamard basis of $\mathbf{R}^n$, and κ_j are independent and uniformly distributed over $\{1,...,n\}$. On the other hand, in the latter case, for $z = x = \psi_1$ we have

$$\mathbf{E}\{(z^T\phi\phi^T x)^2\} = \mathbf{E}\{(\psi_1\phi\phi^T\psi_1)^2\} = n = n\|\psi_1\|_2^4 = n\|x\|_2^2\|z\|_2^2,$$

implying that (36) can only hold with $\chi = O(n)$ in this case.

Besides this, when ϕ is a linear image of a Rademacher vector, i.e. $\phi = A\eta$ where $A \in \mathbf{R}^{m\times n}$ and η has independent components $[\eta]_i \in \{\pm 1\}$ with $\mathrm{Prob}\{[\eta]_i = 1\} = \mathrm{Prob}\{[\eta]_i = -1\} = 1/2$, one has $\Sigma = AA^T$, and $\mathbf{E}\{(\phi^T x)^4\} \le 3\|A^T x\|_2^4$ (cf. the case of sub-Gaussian regressors above). Thus, we have

$$\begin{aligned}\mathbf{E}\{(z^T\phi\phi^T(x-x_*))^2\} &\le \mathbf{E}\{(z^T\phi)^4\}^{1/2}\mathbf{E}\{((x-x_*)^T\phi)^4\}^{1/2} \\ &\le 3z^T\Sigma z\,(x-x_*)^T\Sigma(x-x_*) \le 3\sigma_1(\Sigma)\|z\|_2^2\|x-x_*\|_\Sigma^2\end{aligned}$$

implying (36) with $\chi = 6\underline{\ell}^{-1}\overline{\ell}$. On the other hand, when denoting $\mu = \max_j \|\mathrm{Row}_j(A)\|_2$, we get $\mathrm{Prob}\{\|\phi\|_\infty^4 \ge t\mu\} \le 2ne^{-t^2/2}$ with

$$\mathbf{E}\{\|\phi\|_\infty^2\} \le 2\mu^2[\ln[2n]+1] \;\text{ and }\; \mathbf{E}\{\|\phi\|_\infty^4\} \le 4\mu^4[\ln^2[2n] + 2\ln[2n] + 2].$$

Thus, by (28),

$$\begin{aligned}\zeta^2(x) &\le \left(\overline{\ell}(\mathbf{E}\{\|\phi\|_\infty^4\})^{1/4}(\mathbf{E}\{(\phi^T(x-x_*))^4\})^{1/4} + \sigma(\mathbf{E}\{\|\phi\|_\infty^2\})^{1/2} + \overline{\ell}\sqrt{\upsilon}\|x-x_*\|_\Sigma\right)^2 \\ &\le \left(\overline{\ell}\sqrt{2\sqrt{3}(\ln[2n]+2)}\mu\|x-x_*\|_\Sigma + \sigma\sqrt{2(\ln[2n]+1)}\mu + \overline{\ell}\sqrt{\upsilon}\|x-x_*\|_\Sigma\right)^2 \\ &\le 2\mu^2\overline{\ell}^2\left(\sqrt{2\sqrt{3}(\ln[2n]+2)}+1\right)^2\|x-x_*\|_\Sigma^2 + 4\mu^2(\ln[2n]+1)\sigma^2\end{aligned}$$

which is (8) with $\varkappa\nu \lesssim \mu^2\underline{\ell}^{-1}\overline{\ell}^2\ln n$ and $\varkappa'\varsigma_*^2 \lesssim \mu^2\sigma^2\ln n$.

3. *Scale mixtures:* Let us now assume that

$$\phi \sim \sqrt{Z}\eta, \tag{38}$$

where Z is a scalar a.s. positive random variable, and $\eta \in \mathbf{R}^n$ is independent of Z with covariance matrix $\mathbf{E}\{\eta\eta^T\} = \Sigma_0$. Because

$$\mathbf{E}\{\|\phi\|_\infty^2\} = \mathbf{E}\{Z\}\mathbf{E}\{\|\eta\|_\infty^2\},\;\; \mathbf{E}\{\|\phi\phi^T z\|_\infty^2\} = \mathbf{E}\{Z^2\}\mathbf{E}\{\|\eta\eta^T z\|_\infty^2\}$$

and

$$[\Sigma :=]\ \mathbf{E}\{\phi\phi^T\} = \mathbf{E}\{Z\}\mathbf{E}\{\eta\eta^T\},$$

we conclude that if random vector η satisfies (28) with Σ_0 substituted for Σ and $\mathbf{E}\{Z^2\}$ is finite then a similar bound also holds for ϕ. It is obvious that if η satisfies (36) then

$$\mathbf{E}\big\{(z^T\phi\phi^T x)^2\big\} = \mathbf{E}\{Z^2\}\mathbf{E}\big\{(z^T\eta\eta^T x)^2\big\} \le \frac{\mathbf{E}\{Z^2\}}{\mathbf{E}\{Z\}^2}\chi\|z\|_\Sigma^2\|x\|_\Sigma^2 \le \chi\frac{\mathbf{E}\{Z^2\}}{\mathbf{E}\{Z\}^2}\sigma_1(\Sigma)\|z\|_2^2\|x\|_\Sigma^2,$$

and (36) holds for ϕ with χ for η replaced with $\chi\frac{\mathbf{E}\{Z^2\}}{\mathbf{E}\{Z\}^2}$.

Let us consider the situation where $\eta \sim \mathcal{N}(0, \Sigma_0)$ with positive definite Σ_0. In this case ϕ is referred to as Gaussian scale mixture with a standard example provided by *n-variate t-distributions* $t_n(q, \Sigma_0)$ (multivariate Student distributions with q degrees of freedom, see [34] and references therein). Here, by definition, $t_n(q, \Sigma_0)$ is the distribution of the random vector $\phi = \sqrt{Z}\eta$ with $Z = q/\zeta$, where ζ is the independent of η χ^2-random variable with q degrees of freedom. One can easily see that all one-dimensional projections $e^T\phi$, $\|e\|_2 = 1$, of ϕ are random variables with univariate t_q-distribution. When $\phi_i \sim t_n(q, \Sigma_0)$ with $q > 4$, we have for $\zeta \sim \chi^2_q$

$$\mathbf{E}\left\{\frac{q}{\zeta}\right\} = \frac{q}{q-2}, \quad \mathbf{E}\left\{\frac{q^2}{\zeta^2}\right\} = \frac{3q^2}{(q-2)(q-4)},$$

so that $\Sigma = \frac{q}{q-2}\Sigma_0$, and

$$\varsigma^2(x) \quad \lesssim \quad \bar{\ell}^2\frac{q-2}{q-4}\upsilon\ln[n]\|x - x_*\|_\Sigma + \sigma^2\upsilon\ln n$$

implying (8) with $\varkappa\nu \lesssim \bar{\ell}^2\underline{\ell}^{-1}\upsilon\ln n$, $\varkappa' \lesssim 1$, and $\varsigma_*^2 \lesssim \sigma^2\upsilon\ln n$. Moreover, in this case

$$\mathbf{E}\{(z^T\phi\phi^T x)^2\} = \mathbf{E}\{Z^2\}\mathbf{E}\{z^T\eta\eta^T x)^2\} \le 3\frac{\mathbf{E}\{Z^2\}}{\mathbf{E}\{Z\}^2}\|z\|_\Sigma^2\|x\|_\Sigma^2 \le 9\frac{q-2}{q-4}\sigma_1(\Sigma)\|z\|_2^2\|x\|_\Sigma^2.$$

Another example of Gaussian scale mixture (38) is the *n-variate Laplace distribution* $\mathcal{L}_n(\lambda, \Sigma_0)$ [20] in which Z has exponential distribution with parameter λ. In this case all one-dimensional projections $e^T\phi$, $\|e\|_2 = 1$, of ϕ are Laplace random variables. If $\phi_i \sim \mathcal{L}_n(\lambda, \Sigma_0)$ one has

$$\varsigma^2(x) \lesssim \bar{\ell}^2\upsilon\ln[n]\|x - x_*\|_\Sigma + \sigma^2\upsilon\ln n$$

and

$$\mathbf{E}\{(z^T\phi\phi^T x)^2\} \lesssim \sigma_1(\Sigma)\|z\|_2^2\|x\|_\Sigma^2.$$

3.2 Stochastic Mirror Descent for low-rank matrix recovery

In this section we consider the problem of recovery of matrix $x_* \in \mathbf{R}^{p\times q}$, from independent and identically distributed observations

$$\eta_i = \langle\phi_i, x_*\rangle + \sigma\xi_i, \quad i = 1, 2, ..., N, \tag{39}$$

with $\phi_i \in \mathbf{R}^{p\times q}$ which are random independent over i with covariance operator Σ (defined according to $\Sigma(x) = \mathbf{E}\{\phi\langle\phi, x\rangle\}$). We assume that $\xi_i \in \mathbf{R}$ are mutually independent and independent of ϕ_i with $\mathbf{E}\{\xi_i\} = 0$ and $\mathbf{E}\{\xi_i^2\} \le 1$.

In this application, E is the space of $p \times q$ matrices equipped with the Frobenius scalar product

$$\langle a, b \rangle = \mathrm{Tr}\left(a^T b\right)$$

with the corresponding norm $\|a\|_2 = \langle a, a \rangle^{1/2}$. For the sake of definiteness, we assume that $p \geq q \geq 2$. Our choice for the norm $\|\cdot\|$ is the nuclear norm $\|x\| = \|\sigma(x)\|_1$ where $\sigma(x)$ is the singular spectrum of x, so that the conjugate norm is the spectral norm $\|y\|_* = \|\sigma(y)\|_\infty$. We suppose that

$$\kappa_\Sigma \|x\|_2^2 \leq \langle x, \Sigma(x) \rangle \leq \upsilon \|x\|_2^2 \; \forall x \in \mathbf{R}^{p \times q},$$

with known $\kappa_\Sigma > 0$ and υ, we write $\kappa_\Sigma I \preceq \Sigma \preceq \upsilon I$; for $x \in \mathbf{R}^{p \times q}$ we denote $\|x\|_\Sigma = \sqrt{\langle x, \Sigma(x) \rangle}$. Finally, we assume that matrix x_* is of rank $s \leq \bar{s} \leq q$, and moreover, that we are given a convex and closed subset X of $\mathbf{R}^{p \times q}$ such that $x_* \in X$, along with $R < \infty$ and $x_0 \in X$ satisfying $\|x_* - x_0\| \leq R$.

Consider the Stochastic Optimization problem

$$\min_{x \in X} \left\{ g(x) = \tfrac{1}{2} \mathbf{E}\big\{ \underbrace{(\eta - \langle \phi, x \rangle)^2}_{=: G(x, \omega = [\phi, \eta])} \big\} \right\}. \tag{40}$$

We are to apply SMD algorithm to solve (40) with the proximal setup associated with the nuclear norm with quadratically growing for $q \geq 2$ distance-generating function

$$\vartheta(x) = 2e \ln(2q) \left[\sum_{j=1}^{q} \sigma_j^{1+r}(x) \right]^{\frac{2}{1+r}}, \quad r = \big(12 \ln[2q]\big)^{-1}$$

(here $\sigma_j(x)$ are singular values of x), with the corresponding parameter $\Theta \leq C \ln[2q]$ (cf. [49, Theorem 2.3]). Note that, in the premise of this section,

$$g(x) = \tfrac{1}{2} \mathbf{E}\big\{ (\sigma \xi + \langle \phi, x_* - x \rangle)^2 \big\} = \tfrac{1}{2} (\|x - x_*\|_\Sigma^2 + \sigma^2),$$

with

$$\nabla g(x) = \Sigma(x - x_*) = \mathbf{E}\big\{ \underbrace{\phi(\langle \phi, x - x_* \rangle - \sigma \xi)}_{= \nabla G(x, \omega)} \big\}$$

and

$$\zeta(x, \omega) = \nabla G(x, \omega) - \nabla g(z) = [\phi \langle \phi, x - x_* \rangle - \Sigma(x - x_*)] - \sigma \phi \xi.$$

Let us now consider the case regressors $\phi_i \in \mathbf{R}^{p \times q}$ drawn independently from a *sub-Gaussian ensemble*, $\phi_i \sim \mathrm{Sub}\mathcal{G}(0, S)$ with sub-Gaussian operator S. The latter means that

$$\mathbf{E}\big\{ e^{\langle x, \phi \rangle} \big\} \leq e^{\frac{1}{2} \langle x, S(x) \rangle} \quad \forall x \in \mathbf{R}^{p \times q}$$

with linear positive definite $S(\cdot)$. To show the bound of Theorems 2.1–2.3 in this case we need to verify that relationships (8) and (23) of Assumptions **S1** and **S3** are satisfied. To this end, let us assume that S is "similar" to the covariance operator Σ of ϕ, namely, $S \preceq \mu \Sigma$ with some $\mu < \infty$. This setting covers, for instance, the situation where the entries in the regressors matrix $\phi \in \mathbf{R}^{p \times q}$ are standard Gaussian or Rademacher i.i.d. random variables (in these models, $S = \Sigma$ is the identity, and $g(x) - g_* = \frac{1}{2} \|x - x_*\|_2^2$).

Note that, more generally, when $S \preceq \mu\Sigma$ we have $S \preceq \mu\upsilon I$ with

$$\mathbf{E}\{\|\phi\|_*^4\} \le C^2\mu^2\upsilon^2(p+q)^2,$$

cf. Lemma A.3 of the appendix, and

$$\mathbf{E}\{\langle\phi, x-x_*\rangle^4\} \le 16\langle x-x_*, S(x-x_*)\rangle^2 \le 16\mu^2\|x-x_*\|_\Sigma^2$$

for sub-Gaussian random variable $\langle\phi, x-x_*\rangle \sim \mathrm{Sub}\mathcal{G}(0, \langle x-x_*, S(x-x_*)\rangle)$. Therefore,

$$\begin{aligned}\mathbf{E}\big\{\|\phi\langle\phi, x-x_*\rangle - \Sigma(x-x_*)\|_*^2\big\} &\le 2\mathbf{E}\big\{\|\phi\langle\phi, x-x_*\rangle\|_*^2\big\} + 2\upsilon\|x-x_*\|_\Sigma^2\\ &\le 2\mathbf{E}\big\{\|\phi\|_*^4\big\}^{1/2}\mathbf{E}\big\{\langle\phi, x-x_*\rangle^4\big\}^{1/2} + 2\upsilon\|x-x_*\|_\Sigma^2\\ &\le 8C\mu^2(p+q)\upsilon\|x-x_*\|_\Sigma^2 + 2\upsilon\|x-x_*\|_\Sigma^2.\end{aligned}$$

Taking into account that $\nu = \mathbf{E}\{\|\phi\|_*^2\} \le C\mu\upsilon(p+q)$ in this case, we have

$$\begin{aligned}\varsigma^2(x) &= \mathbf{E}\big\{\|\zeta(x,\omega)\|_*^2\big\} \le 2\mathbf{E}\big\{\|\phi\langle\phi, x-x_*\rangle - \Sigma(x-x_*)\|_*^2\big\} + 2\sigma^2\mathbf{E}\big\{\|\phi\|_*^2\big\}\\ &\le 8(4C\mu^2(p+q)+1)\upsilon[g(x)-g_*] + 2\underbrace{C\mu\upsilon(p+q)\sigma^2}_{=\varsigma_*^2}\end{aligned}$$

implying (8) with $\varkappa \lesssim \mu$ and $\varkappa' \lesssim 1$.

Similarly, we estimate $\forall x \in X$, $z \in \mathbf{R}^{p\times q}$

$$\mathbf{E}\left\{\langle\phi, z\rangle^2\langle\phi, x\rangle^2\right\} \le \mathbf{E}\left\{\langle z, \phi\rangle^4\right\}^{1/2}\mathbf{E}\left\{\langle\phi, x\rangle^4\right\}^{1/2} \le 16\langle z, S(z)\rangle\langle x, S(x)\rangle \le 16\mu^2\upsilon\|z\|_2^2\|x\|_\Sigma^2,$$

so that

$$\begin{aligned}\mathbf{E}\left\{\langle z, \zeta(x,\omega)\rangle^2\right\} &= \mathbf{E}\left\{\langle z, \phi\langle\phi, x-x_*\rangle - \Sigma(x-x_*) - \sigma\phi\xi\rangle^2\right\}\\ &= \mathbf{E}\left\{\big(\langle z, \phi\rangle\langle\phi, x-x_*\rangle - \langle z, \Sigma(x-x_*)\rangle\big)^2\right\} + \sigma^2\mathbf{E}\{\xi^2\langle z, \phi\rangle^2\}\\ &\le \mathbf{E}\left\{\langle z, \phi\rangle^2\langle\phi, x-x_*\rangle^2\right\} + \sigma^2\upsilon\|z\|_2^2\\ &\le 16\mu^2\upsilon[g(x)-g_*]\|z\|_2^2 + \sigma^2\upsilon\|z\|_2^2\end{aligned}$$

implying the bound (23) with $\chi \lesssim \mu(p+q)^{-1}$ and $\chi' \lesssim \mu^{-1}(p+q)^{-1}$. When substituting the above bounds for problem parameters into statements of Theorems 2.1–2.3 we obtain the following statement summarizing the properties of the approximate solutions by the SMD-SR algorithm utilizing observations (39); the corresponding risks are defined in (30)–(33).

Proposition 3.3 *In the situation of this section,*

(i) let the sample size N satisfy

$$N \ge \alpha\left\lceil\frac{\mu^2\upsilon(p+q)\bar{s}\ln q}{\kappa_\Sigma}\right\rceil$$

for an appropriate absolute α, implying that at least one preliminary stage of Algorithm 1 is completed. Then there is an absolute $c > 0$ such that approximate solutions $\widehat{x}_N$ and $\widehat{y}_N$ produced by the algorithm satisfy

$$\begin{aligned}\mathrm{Risk}_{\|\cdot\|}(\widehat{y}_N|X) &\le 2\sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2}(\widehat{x}_N|X) \lesssim R\exp\left\{-\frac{cN\kappa_\Sigma}{\mu^2\upsilon(p+q)\bar{s}\ln q}\right\} + \frac{\sigma\bar{s}}{\kappa_\Sigma}\sqrt{\frac{\mu\upsilon(p+q)\ln q}{N}},\\ \mathrm{Risk}_g(\widehat{x}_N|X) &\lesssim \frac{\kappa_\Sigma R^2}{\bar{s}}\exp\left\{-\frac{cN\kappa_\Sigma}{\mu^2\upsilon(p+q)\bar{s}\ln q}\right\} + \frac{\sigma^2\mu\upsilon(p+q)\bar{s}\ln q}{\kappa_\Sigma N}.\end{aligned}$$

(ii) Furthermore, when observation size satisfies

$$N \geq \alpha' \left[\frac{\mu^2 \upsilon (p+q) \bar{s} \ln[1/\epsilon] \ln q}{\kappa_\Sigma} \right]$$

with large enough α', $(1-\epsilon)$-reliable solutions $\widehat{y}_{N,1-\epsilon}$ and $\widehat{x}_{N,1-\epsilon}$ defined in Section 2.4 satisfy for some $c' > 0$

$$\begin{aligned} \mathrm{Risk}_{\|\cdot\|,\epsilon}(\widehat{y}_{N,1-\epsilon}|X) &\leq \sqrt{2s}\mathrm{Risk}_{\|\cdot\|_{2,\epsilon}}(\widehat{y}_{N,1-\epsilon}|X) \leq 2\sqrt{2s}\mathrm{Risk}_{\|\cdot\|_{2,\epsilon}}(\widehat{x}_{N,1-\epsilon}|X) \\ &\lesssim R\exp\left\{-\frac{c'N\kappa_\Sigma}{\mu^2\upsilon(p+q)\bar{s}\ln[1/\epsilon]\ln q}\right\} + \frac{\sigma\bar{s}}{\kappa_\Sigma}\sqrt{\frac{\mu\upsilon(p+q)\ln[1/\epsilon]\ln q}{N}}, \end{aligned} \tag{41}$$

with solutions $\widehat{x}'_{N,1-\epsilon}$, $\widehat{x}''_{N,1-\epsilon}$ and $\widehat{y}'_{N,1-\epsilon}$, $\widehat{y}''_{N,1-\epsilon}$ verifying analogous bounds. Finally, the following bound holds for the aggregated solution $\overline{x}_{2N,1-\epsilon}$ (with $K = N$) by Algorithm 2:

$$\mathrm{Risk}_{g,\epsilon}(\overline{x}_{2N,1-\epsilon}|X) \lesssim \frac{\kappa_\Sigma R^2}{\bar{s}}\exp\left\{-\frac{c'N\kappa_\Sigma}{\mu^2\upsilon(p+q)\bar{s}\ln[1/\epsilon]\ln q}\right\} + \frac{\sigma^2\mu\upsilon(p+q)\bar{s}\ln[1/\epsilon]\ln q}{\kappa_\Sigma N}.$$

Remark. Let us now compare the bounds of the proposition to available accuracy estimates for low rank matrix recovery. Notice first, that when assuming that $\mu \lesssim 1$ the bounds of the proposition hold if (the upper bound on unknown) signal rank $\bar{s}$ satisfies

$$\bar{s} \lesssim \frac{N\kappa_\Sigma}{(p+q)\upsilon\ln[1/\epsilon]\ln q}.$$

The above condition is essentially the same, up to logarithmic in $1/\epsilon$ factor, as the best condition on rank of the signal to be recovered under which the recovery is exact in the case of exact—noiseless—observation [11, 53]. The risk bounds of Proposition 3.3 can be compared to the corresponding accuracy bounds for recovery $\widehat{x}_{N,\mathrm{Lasso}}$ by Lasso with nuclear norm penalization, as in [33, 44]. For instance, when regressors ϕ_i have i.i.d. $\mathcal{N}(0,1)$ entries they state (cf. [44, Corollary 5]) that the $\|\cdot\|_{2,\epsilon}$-risk of the recovery satisfies the bound

$$\mathrm{Risk}_{\|\cdot\|_{2,\epsilon}}(\widehat{x}_{N,\mathrm{Lasso}}|X) \lesssim \frac{\sigma^2 r(p+q)}{N}$$

for $\epsilon \geq \exp\{-(p+q)\}$. Observe that the above bound coincides, up to logarithmic in q and $1/\epsilon$ factors with the second—asymptotic—term in the bound (41). This result is all the more surprising if we recall that its validity is not limited to sub-Gaussian regressors—what we need in fact is the bound (cf. the remark after Proposition 3.2)

$$\mathbf{E}\{\|\phi\langle\phi, z\rangle\|_*^2\} \lesssim (p+q)\|x - x_*\|_\Sigma^2. \tag{42}$$

For instance, one straightforwardly verifies that the latter bound holds, for instance, in the case where regressor ϕ is a scale mixtures of matrices satisfying (42) (e.g., scale mixture of sub-Gaussian matrices).

4 Numerical illustration

We present results of a preliminary simulation study illustrating performance of the SMD-SR algorithm.

Experimental setting. We present results of simulated experiments of sparse linear regression (25) with linear activation $\mathfrak{u}(t) = t$ and i.i.d. random (ϕ_i, ξ_i) in the setting $N \leq n$ with $(n, s) = (100\,000, 50)$. In our experiments, covariance matrix Σ of regressors is diagonal with diagonal entries $\Sigma_{11} \leq \Sigma_{22} \leq \cdots \leq \Sigma_{nn}$ evenly spaced over $[\kappa_\Sigma, \nu]$, parameters (κ_Σ, ν) being specific for each experiment. The indices of nonvanishing components of the optimal solution x_* are evenly spaced in $[1, n]$ with the non-zero entries being sampled from the standard Gaussian distribution. The number s of nonzero components of x_* and the value κ_Σ are assumed to be known.

We compare the performance of the SMD-SR procedure to that of the "vanilla" non-Euclidean SMD algorithm utilizing the same proximal setup when solving stochastic optimization problem (27). Another contender is the coordinate descent algorithm (CDA) of the Python package `sklearn` solving the Lasso problem

$$\min_{x \in \mathbf{R}^n} \left\{ \frac{1}{2N} \sum_{i=1}^{N} [\eta_i - \phi_i^T x]^2 + \lambda \|x\|_1 \right\} \tag{43}$$

with the "theoretically optimal" choice $\lambda = 2\sigma\sqrt{\frac{2\ln n}{N}}$ of the penalty parameter (cf. [3, 33]).

Parameter setting for SMD-SR. As it is often the case, the theoretical choice of algorithm parameters as given in Sections 2.3 and 3.1 is too conservative in practice. We give a brief overview of the workarounds used in our simulations.

- We use stages of fixed length and mini-batches of exponentially increasing size during the asymptotic phase of the method, cf. [35, Section 4.5]. This allows to significantly accelerate computations at the asymptotic regime alleviating the computational burden of prox-evaluations.
- We use variable stepsize parameters $\beta_i = \beta_0 \|\phi_i\|_\infty^2$ with constant $\beta_0 = 1.0$ both for SMD-SR and SMD. This choice of β_0 corresponds to the condition $\beta_0 \geq \nu$ but neglects the constants factors arising in the theoretical analysis. In order to compute the current approximate solution, the estimates of the SMD algorithm are then weighted with the corresponding β_i.
- The number of steps m_0 to be performed by the SMD algorithm on each stage is set to $m_0 = \lceil (1/2) s\nu(\ln[n] + 1) \rceil$, which corresponds to (15) in the case of $\kappa_\Sigma = 1.0$.
- In our simulations, we utilize the CUSUM test for monitoring a change detection (see, e.g., [39, 51]) to decide upon switching from preliminary ("linear trend") to asymptotic phase ("sublinear trend") of the algorithm; however, we perform at least 4 preliminary stages.

Experimental results. We present results of two series of experiments, experiments in each series corresponding to 4 combinations of parameters κ_Σ and σ with $\kappa_\Sigma \in \{0.1, 1.0\}$ and $\sigma \in \{0.001, 0.1\}$; we run 20 simulations for each parameter combination. In the figures below, for each "contender" we plot the median value of the prediction error $\|\widehat{x}_t - x_*\|_\Sigma$ as a function of $t = 1, ..., N$ along with the tubes of 25% and 75% quantiles.

In the first series of simulations, noises (ξ_i) are standard Gaussian, and regressors (ϕ_i) are normally distributed with zero mean and covariance matrix Σ. The results for the first series are presented in Figures 1 and 2. Plots in Figure 1 illustrate the improvement by the SMD-SR

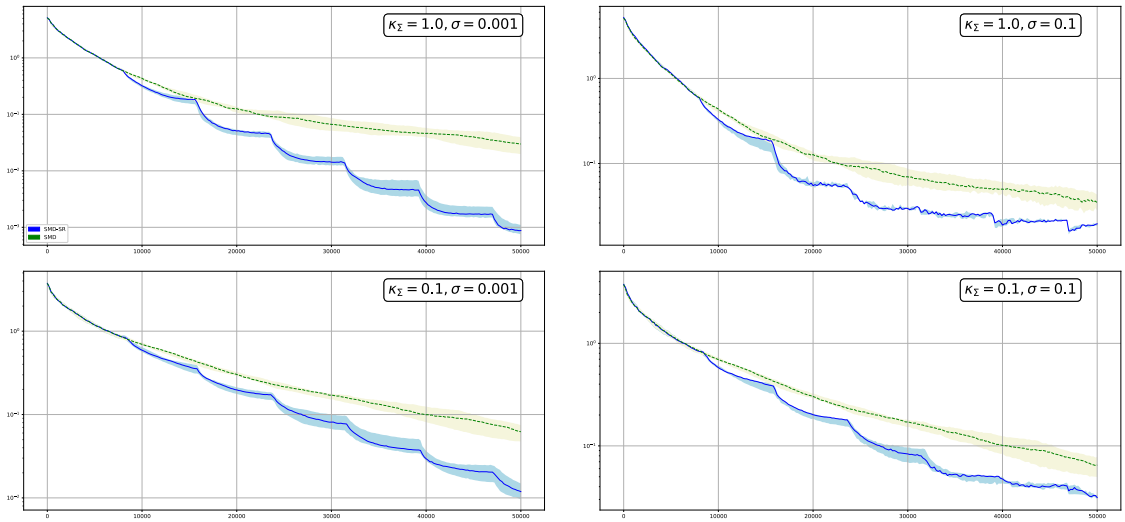

Figure 1: Comparison of SMD-SR (solid line) and SMD (dashed line) in the Gaussian setting; $(n, s) = (100\,000, 50)$.

procedure over the plain SMD algorithm in the considered settings. The acceleration of the initial error convergence is clearly seen on the plots for $\sigma = 0.001$.

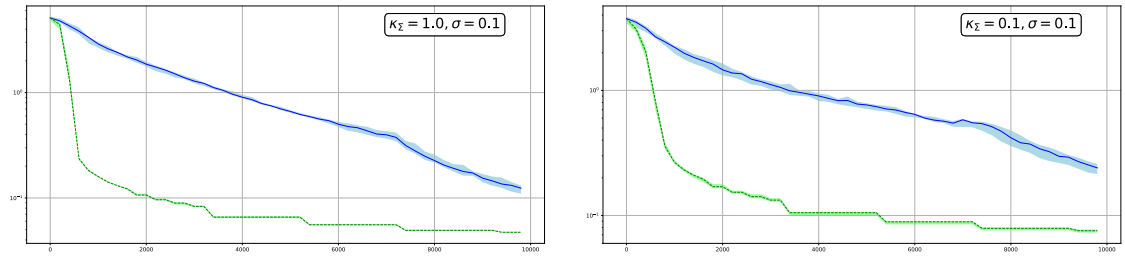

Figure 2: Comparison of SMD-SR (solid line) and Lasso by CDA (dashed line) in the Gaussian setting; $(n, s) = (50\,000, 50)$.

Results of a comparison with the CDA Lasso implementation of in the case of $\sigma = 0.1$ are given in Figure 2. Because of the memory limitations of the CDA, we present the results of simulations for $(n, s) = (50\,000, 50)$ and $N \leq 10\,000$. The CDA is restarted for different sizes of the observation sample, each time the number of iterations of the algorithm is limited to 30 000. While Lasso estimate outperforms the SMD-SR for smaller observation samples, the statistical performance of the proposed algorithm appears to be competitive for large N.

Similar results were obtained in the experiments with other types of distributions of ϕ_i and ξ_i. For instance, in Figure 3 we present the results of simulation utilizing Student's t_4-distribution (i.e., multivariate Student distribution with 4 degrees of freedom, cf., e.g., [34]) of noises and regressors.

Acknowledgment

This work was supported by Multidisciplinary Institute in Artificial intelligence MIAI @ Grenoble Alpes (ANR-19-P3IA-0003).

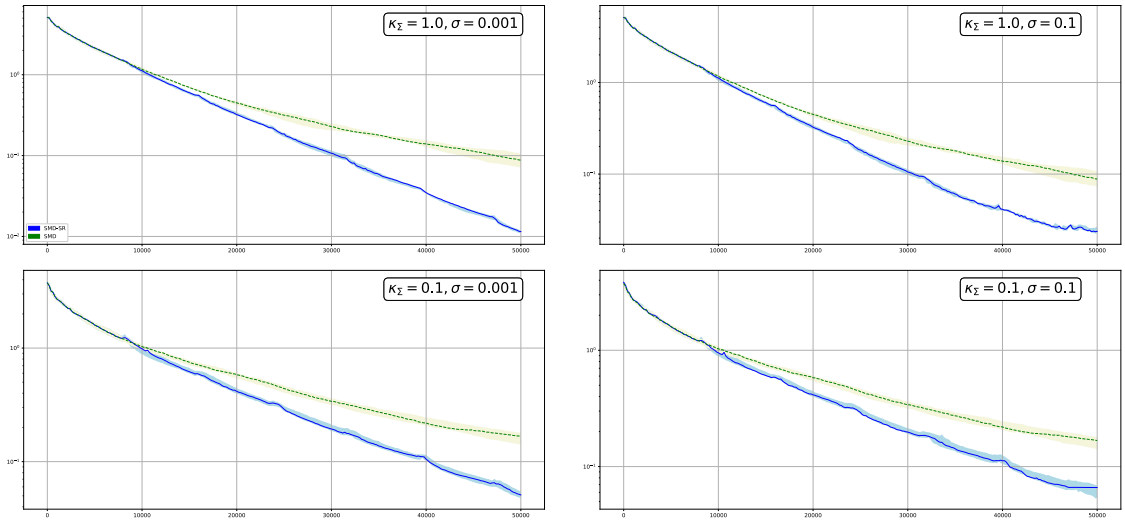

Figure 3: Comparison of SMD-SR (solid line) and SMD (dashed line) in the case of Student t_4 noise distribution; $(n, s) = (100000, 50)$.

A Proofs

A.1 Proof of Proposition 2.1

We start with a technical result on the SMD algorithm which we formulate in a more general setting of composite minimization. Specifically, assume that we aim at solving the problem

$$\min_{x\in X} [f(x) = \mathbf{E}\{G(x,\omega)\} + h(x)], \tag{44}$$

where X and G are as in Section 2.1 and h is convex and continuous. We consider a more general *composite proximal mapping* [48, 49] for $\zeta \in E$, $x, x_0 \in X$, and $\beta > 0$ we define

$$\begin{array}{rcl} \mathrm{Prox}_\beta(\zeta, x; x_0) & := & \mathrm{argmin}_{z\in X}\left\{\langle \zeta, z\rangle + h(z) + \beta V_{x_0}(x, z)\right\} \\ & = & \mathrm{argmin}_{z\in X}\left\{\langle \zeta - \beta\nabla\vartheta(x - x_0), z\rangle + h(z) + \beta\vartheta(z - x_0)\right\} \end{array} \tag{45}$$

and consider for $i = 1, 2, \ldots$ Stochastic Mirror Descent recursion (12). Same as before, the approximate solution after N iterations of the algorithm is defined as weighted average of x_i's according to (13). Obviously, to come back to the situation of Section 2.2 it suffices to put $h(x) \equiv 0$. To alleviate notation we denote $V(x, z) = V_{x_0}(x, z)$; we also denote

$$\zeta_i = \nabla G(x_{i-1}, \omega_i) - \nabla g(x_{i-1})$$

and

$$\varepsilon(x^N, z) = \sum_{i=1}^{N} \beta_{i-1}^{-1}[\langle \nabla g(x_{i-1}), x_i - z\rangle + h(x_i) - h(z)] + \tfrac{1}{2}V(x_{i-1}, x_i), \tag{46}$$

with $x^N = (x_0, \ldots, x_N)$. In the sequel we use the following well known result which we prove below for the sake of completeness.

Proposition A.1 *In the situation of this section, let $\beta_i \geq 2\mathcal{L}$ for all $i = 0, 1, ...,$ and let $\widehat{x}_N$ be defined in (13), where x_i are iterations (12). Then for any $z \in X$ we have*

$$\begin{aligned}\left[\sum_{i=1}^N \beta_{i-1}^{-1}\right][f(\widehat{x}_N) - f(z)] &\leq \sum_{i=1}^N \beta_{i-1}^{-1}[f(x_i) - f(z)] \leq \varepsilon(x^N, z)\\ &\leq V(x_0, z) - V(x_N, z) + \sum_{i=1}^N \Big[\frac{\langle \zeta_i, z - x_{i-1}\rangle}{\beta_{i-1}} + \frac{\|\zeta_i\|_*^2}{\beta_{i-1}^2}\Big] \qquad (47)\\ &\leq 2V(x_0, z) + \sum_{i=1}^N \Big[\frac{\langle \zeta_i, z_{i-1} - x_{i-1}\rangle}{\beta_{i-1}} + \frac{3}{2}\frac{\|\zeta_i\|_*^2}{\beta_{i-1}^2}\Big], \qquad (48)\end{aligned}$$

where z_i is a random vector with values in X depending only on $x_0, \zeta_1, \ldots, \zeta_i$.

Proof of Proposition A.1. **1^o.** Let $x_0, \ldots, x_N$ be some points of X; let

$$\varepsilon_{i+1}(z) := \langle \nabla g(x_i), x_{i+1} - z\rangle + \langle h'(x_{i+1}), x_{i+1} - z\rangle + \mathcal{L}V(x_i, x_{i+1})$$

(here $h'(x)$ stands for a subgradient of h at x). Note that $V(x, z) \geq \frac{1}{2}\|x - z\|^2$ due to the strong convexity of $V(x, \cdot)$. Thus, by convexity of g and h and the Lipschitz continuity of ∇g we get for any $z \in X$

$$\begin{aligned}f(x_{i+1}) - f(z) &= [g(x_{i+1}) - g(z)] + [h(x_{i+1}) - h(z)]\\ &= [g(x_{i+1}) - g(x_i)] + [g(x_i) - g(z)] + [h(x_{i+1}) - h(z)]\\ &\leq [\langle \nabla g(x_i), x_{i+1} - x_i\rangle + \mathcal{L}V(x_i, x_{i+1})] + \langle \nabla g(x_i), x_i - z\rangle + h(x_{i+1}) - h(z)\\ &\leq \langle \nabla g(x_i), x_{i+1} - z\rangle + \langle h'(x_{i+1}), x_{i+1} - z\rangle + \mathcal{L}V(x_i, x_{i+1}) = \varepsilon_{i+1}(z);\end{aligned}$$

i.e., the following inequality holds for any $z \in X$:

$$f(x_{i+1}) - f(z) \leq \varepsilon_{i+1}(z). \qquad (49)$$

2^o. Let us first prove inequality (47). The optimality condition for x_{i+1} in (45) implies (cf. Lemma A.1 of [49]) that there is $h'(x_{i+1}) \in \partial h(x_{i+1})$ such that

$$\langle \nabla G(x_i, \omega_{i+1}) + h'(x_{i+1}) + \beta_i\langle[\nabla\vartheta(x_{i+1}) - \nabla\vartheta(x_i)], z - x_{i+1}\rangle \geq 0, \ \ \forall\, z \in X,$$

or, equivalently,

$$\begin{aligned}&\langle \nabla G(x_i, \omega_{i+1}) + h'(x_{i+1}), x_{i+1} - z\rangle \leq \beta_i\langle \nabla\vartheta(x_{i+1}) - \nabla\vartheta(x_i), z - x_{i+1}\rangle\\ &\quad = \beta_i\langle \nabla V_{x_{i+1}}(x_i, x_{i+1}), z - x_{i+1}\rangle = \beta_i[V(x_i, z) - V(x_{i+1}, z) - V(x_i, x_{i+1})], \ \ \forall\, z \in X\end{aligned}$$

where the concluding equality follows from the following remarkable identity (see, for instance, [18]): for any u, u' and $w \in X$

$$\langle \nabla_{u'} V(u, u'), w - u'\rangle = V(u, w) - V(u', w) - V(u, u').$$

This results in

$$\begin{aligned}\langle \nabla g(x_i), x_{i+1} - z\rangle + \langle h'(x_{i+1}), x_{i+1} - z\rangle &\leq \beta_i[V(x_i, z) - V(x_{i+1}, z) - V(x_i, x_{i+1})]\\ &\quad -\langle \zeta_{i+1}, x_{i+1} - z\rangle. \qquad (50)\end{aligned}$$

It follows from (49) and condition $\beta_i \geq 2\mathcal{L}$ that

$$f(x_{i+1}) - f(z) \leq \varepsilon_{i+1}(z) \leq \langle \nabla g(x_i), x_{i+1} - z\rangle + \langle h'(x_{i+1}), x_{i+1} - z\rangle + \frac{\beta_i}{2} V(x_i, x_{i+1}).$$

Together with (50), this inequality implies

$$\varepsilon_{i+1}(z) \leq \beta_i[V(x_i, z) - V(x_{i+1}, z) - \tfrac{1}{2}V(x_i, x_{i+1})] - \langle \zeta_{i+1}, x_{i+1} - z\rangle.$$

On the other hand, due to the strong convexity of $V(x, \cdot)$ we have

$$\begin{aligned}\langle \zeta_{i+1}, z - x_{i+1}\rangle - \frac{\beta_i}{2} V(x_i, x_{i+1}) &= \langle \zeta_{i+1}, z - x_i\rangle + \langle \zeta_{i+1}, x_i - x_{i+1}\rangle - \frac{\beta_i}{2} V(x_i, x_{i+1}) \\ &\leq \langle \zeta_{i+1}, z - x_i\rangle + \frac{\|\zeta_{i+1}\|_*^2}{\beta_i}.\end{aligned}$$

Combining these inequalities, we obtain

$$f(x_{i+1}) - f(z) \leq \varepsilon_{i+1}(z) \leq \beta_i[V(x_i, z) - V(x_{i+1}, z)] - \langle \zeta_{i+1}, x_i - z\rangle + \frac{\|\zeta_{i+1}\|_*^2}{\beta_i} \tag{51}$$

for all $z \in X$. Dividing (51) by β_i and taking the sum over i from 0 to $N-1$ we obtain (47).

3^o. We now prove the bound (48). Applying Lemma 6.1 of [45] with $z_0 = x_0$ we get

$$\forall z \in X, \qquad \sum_{i=1}^N \beta_{i-1}^{-1}\langle \zeta_i, z - z_{i-1}\rangle \leq V(x_0, z) + \tfrac{1}{2}\sum_{i=1}^N \beta_{i-1}^{-2}\|\zeta_i\|_*^2, \tag{52}$$

where $z_i = \operatorname{argmin}_{z\in X}\big\{ - \beta_{i-1}^{-1}\langle \zeta_i, z\rangle + V(z_{i-1}, z)\big\}$ depend only on $z_0, \zeta_1, \ldots, \zeta_i$. Further,

$$\begin{aligned}\sum_{i=1}^N \beta_{i-1}^{-1}\langle \zeta_i, z - x_{i-1}\rangle &= \sum_{i=1}^N \beta_{i-1}^{-1}[\langle \zeta_i, z_{i-1} - x_{i-1}\rangle + \langle \zeta_i, z - z_{i-1}\rangle] \\ &\leq V(x_0, z) + \sum_{i=1}^N \beta_{i-1}^{-1}\langle \zeta_i, z_{i-1} - x_{i-1}\rangle + \tfrac{1}{2}\beta_{i-1}^{-2}\|\zeta_i\|_*^2.\end{aligned}$$

Combining this inequality with (47) we arrive at (48). □

Proof of Proposition 2.1. Note that, by definition, $\nu \geq \mathcal{L}$ and $\varkappa \geq 1$, thus, Proposition A.1 can be applied to the corresponding SMD recursion. When applying recursively bound (47) of the proposition with $z = x_*$ and $h(x) \equiv 0$ we conclude that $\mathbf{E}\{V_{x_0}(x_i, x_*)\}$ is finite along with $\mathbf{E}\{\|x_i - x_*\|^2\}$, and so $\mathbf{E}\{\langle \zeta_{i+1}, x_i - x_*\rangle\} = 0$. Thus, after taking expectation we obtain

$$\begin{aligned}\sum_{i=1}^m [\mathbf{E}\{g(x_i)\} - g_*] &\leq \beta \mathbf{E}\{V_{x_0}(x_0, x_*) - V_{x_0}(x_m, x_*)\} + \beta^{-1}\sum_{i=1}^m \mathbf{E}\{\|\zeta_i\|_*^2\} \\ &\leq \mathbf{E}\{V_{x_0}(x_0, x_*) - V_{x_0}(x_m, x_*)\} \\ &\quad + \beta^{-1}\sum_{i=1}^m \big(\varkappa\nu[\mathbf{E}\{g(x_{i-1}) - \langle \nabla g(x_*), x_{i-1} - x_*\rangle\} - g_*] + \varkappa' \varsigma_*^2\big),\end{aligned}$$

which, thanks to convexity of g, leads to

$$\begin{aligned}&\left[1 - \frac{\varkappa\nu}{\beta}\right]\sum_{i=1}^m [\mathbf{E}\{g(x_i)\} - g_*] + \beta \mathbf{E}\{V_{x_0}(x_m, x_*)\} \\ &\leq \beta \mathbf{E}\{V_{x_0}(x_0, x_*)\} + \frac{\varkappa\nu}{\beta}[\mathbf{E}\{g(x_0) - \langle \nabla g(x_*), x_0 - x_*\rangle\} - g_*] + \frac{m\varkappa'\varsigma_*^2}{\beta}.\end{aligned}$$

Because, due to convexity of g, $g(\widehat{x}_m) \le \frac{1}{m}\sum_{i=1}^m g(x_i)$ and

$$\mathbf{E}\{g(x_0) - \langle \nabla g(x_*), x_0 - x_*\rangle\} - g_* \le \tfrac{1}{2}\nu\mathbf{E}\{\|x_0 - x_*\|^2\} \le \tfrac{1}{2}\nu R^2$$

we conclude that when $\beta \ge 2\varkappa\nu$

$$\mathbf{E}\{g(\widehat{x}_m)\} - g_* \le \frac{2R^2}{m}\left(\Theta\beta + \frac{\varkappa\nu^2}{2\beta}\right) + \frac{2\varkappa'\varsigma_*^2}{\beta}$$

which is (14). □

A.2 Proof of Theorem 2.1

We start with the following straightforward result:

Lemma A.1 *Let $x_* \in X \subset E$ be s-sparse, $x \in X$, and let $x_s = \mathrm{sparse}(x)$—an optimal solution to (9). We have*

$$\|x_s - x_*\| \le \sqrt{2s}\|x_s - x_*\|_2 \le 2\sqrt{2s}\|x - x_*\|_2. \tag{53}$$

Proof. Indeed, we have

$$\|x_s - x_*\|_2 \le \|x_s - x\|_2 + \|x - x_*\|_2 \le 2\|x - x_*\|_2$$

(recall that x_* is s-sparse). Because $x_s - x_*$ is $2s$-sparse we have by Assumption **S2**

$$\|x_s - x_*\| \le \sqrt{2s}\|x_s - x_*\|_2 \le 2\sqrt{2s}\|x - x_*\|_2.$$ □

Proof of the theorem relies upon the following characterization of the properties of approximate solutions y_k, x_k, x'_k and y'_k.

Proposition A.2 *Under the premise of Theorem 2.1,*

(i) after k preliminary stages of the algorithm one has

$$\mathbf{E}\{\|y_k - x_*\|^2\} \le 2s\mathbf{E}\{\|y_k - x_*\|_2^2\} \le 2^{-k}R^2 + 32\frac{\varsigma_*^2\bar{s}\varkappa'}{\underline{\kappa}\nu\varkappa}, \tag{54}$$

$$\mathbf{E}\{g(\widehat{x}_{m_0}(y_{k-1},\beta))\} - g_* \le 2^{-k-4}\frac{\underline{\kappa}R_0^2}{\bar{s}} + \frac{2\varkappa'\varsigma_*^2}{\varkappa\nu}. \tag{55}$$

In particular, upon completion of $K = \overline{K}$ preliminary stages approximate solutions $\widehat{x}^{(1)}$ and $\widehat{y}^{(1)}$ satisfy

$$\mathbf{E}\{\|\widehat{y}^{(1)} - x_*\|^2\} \le 2s\mathbf{E}\{\|\widehat{y}^{(1)} - x_*\|_2^2\} \le 64\frac{\varsigma_*^2\bar{s}\varkappa'}{\underline{\kappa}\nu\varkappa}, \tag{56}$$

$$\mathbf{E}\{g(\widehat{x}^{(1)})\} - g_* \le \frac{4\varkappa'\varsigma_*^2}{\varkappa\nu}. \tag{57}$$

(ii) Suppose that at least one asymptotic stage is complete. Let $r_k^2 = 2^{-k}r_0^2$ where $r_0^2 = 64\frac{\varsigma_*^2\bar{s}\varkappa'}{\underline{\kappa}\nu\varkappa}$. Then after k stages of the asymptotic phase one has

$$\mathbf{E}\{\|y'_k - x_*\|^2\} \le 2s\mathbf{E}\{\|y'_k - x_*\|_2^2\} \le r_k^2 = 2^{-k}r_0^2, \tag{58}$$

$$\mathbf{E}\{g(\widehat{x}_{m_k}(y'_{k-1},\beta))\} - g_* \le \frac{4\varsigma_*^2\varkappa'}{\beta_k} \le 2^{-k+2}\frac{\varsigma_*^2\varkappa'}{\varkappa\nu}. \tag{59}$$

Proof of the proposition. 1^o. We first show that under the premise of the proposition the following relationship holds for $1 \le k \le K$:

$$\mathbf{E}\{\|y_k - x_*\|^2\} \le R_k^2 := \tfrac{1}{2}R_{k-1}^2 + \frac{16\varsigma_*\bar{s}\varkappa'}{\underline{\kappa}\nu\varkappa}, \quad R_0 = R. \tag{60}$$

Obviously, (60) implies (54) for all $1 \le k \le K$. Observe that (60) clearly holds for $k = 1$. Let us now perform the recursive step $k-1 \to k$. Indeed, bound (14) of Proposition 2.1 implies that after m_0 iterations of the SMD with the stepsize parameter satisfying (15) and initial condition x_0 such that $\mathbf{E}\{\|x_0 - x_*\|^2\} \le R_{k-1}$ one has

$$\begin{aligned}\mathbf{E}\{g(\widehat{x}_{m_0})\} - g_* &\le \frac{2}{m_0}\left[2\Theta\varkappa\nu + \frac{\nu}{4}\right]R_{k-1}^2 + \frac{\varkappa'\varsigma_*^2}{\varkappa\nu} \\ &\le \frac{[8\Theta\varkappa+1]\nu}{2m_0}R_{k-1}^2 + \frac{\varkappa'\varsigma_*^2}{\varkappa\nu}.\end{aligned} \tag{61}$$

Note that when $m_0 \ge 16\underline{\kappa}^{-1}\bar{s}(8\Theta\varkappa+1)\nu$ we have

$$\frac{8\bar{s}}{\underline{\kappa}}\frac{[8\Theta\varkappa+1]\nu}{m_0} \le \tfrac{1}{2}.$$

Therefore, when utilizing the bound (53) of Lemma A.1 we get

$$\begin{aligned}\mathbf{E}\{\|y_k - x_*\|^2\} &\le 2\bar{s}\mathbf{E}\{\|y_k - x_*\|^2\} \le 8\bar{s}\mathbf{E}\{\|\widehat{x}_{m_0} - x_*\|_2^2\} \le \frac{16\bar{s}}{\underline{\kappa}}[\mathbf{E}\{g(\widehat{x}_{m_0})\} - g_*] \\ &\le \frac{16\bar{s}}{\underline{\kappa}}\left(\frac{[8\Theta\varkappa+1]\nu}{2m_0}R_{k-1}^2 + \frac{\varkappa'\varsigma_*^2}{\varkappa\nu}\right) \le R_k^2 := \tfrac{1}{2}R_{k-1}^2 + \frac{16\varsigma_*^2\bar{s}\varkappa'}{\underline{\kappa}\varkappa\nu}\end{aligned}$$

which is (60). Finally, when using (61) along with (54) we obtain

$$\mathbf{E}\{g(\widehat{x}_{m_0}(y_{k-1},\beta))\} - g_* \le \frac{\underline{\kappa}R_{k-1}^2}{32\bar{s}} + \frac{\varkappa'\varsigma_*^2}{\varkappa\nu} \le 2^{-k-4}\frac{\underline{\kappa}R_0^2}{\bar{s}} + \frac{2\varkappa'\varsigma_*^2}{\varkappa\nu}$$

what implies (55). Now, (56) and (57) follow straightforwardly by applying (54) and (55) with $K = \overline{K}$.

2^o. Let us prove (58). Recall that at the beginning of the first stage of the second phase we have $\mathbf{E}\{\|\bar{y}_0 - x_*\|\} \le r_0^2$. Now, let us do the recursive step, i.e., assume that (58) holds for some $0 \le k < K'$, and let us show that it holds for $k+1$. Because $\Theta \ge 1$ and $\varkappa \ge 1$ we have $\beta_k^2 \ge \frac{\varkappa\nu^2}{2\Theta}$, $k = 1, ...,$ and, by (14),

$$\begin{aligned}\mathbf{E}\{g(\widehat{x}_{m_k}(y'_{k-1},\beta_k))\} - g_* &\le \frac{2r_{k-1}^2}{m_k}\left(\Theta\beta_k + \frac{\varkappa\nu^2}{2\beta_k}\right) + \frac{2\varkappa'\varsigma_*^2}{\beta_k} \le \frac{4\Theta\beta_k r_{k-1}^2}{m_k} + \frac{2\varkappa'\varsigma_*^2}{\beta_k} \\ &\le 2^{-k}\frac{r_0^2\underline{\kappa}}{64\bar{s}} + 2^{1-k}\frac{\varkappa'\varsigma_*^2}{\varkappa\nu} \le 2^{-k}\frac{r_0^2\underline{\kappa}}{16\bar{s}} \le 2^{-k+2}\frac{\varkappa'\varsigma_*^2}{\varkappa\nu}.\end{aligned} \tag{62}$$

Observe that

$$\mathbf{E}\{\|x_{m_k}(y'_{k-1},\beta_k) - x_*\|_2^2\} \le \frac{2}{\underline{\kappa}}[\mathbf{E}\{g(\widehat{x}_{m_k}(y'_{k-1},\beta_k))\} - g_*] \le 2^{-k}\frac{r_0^2}{8\bar{s}},$$

so that by Lemma A.1

$$\mathbf{E}\{\|y'_k - x_*\|^2\} \le 8s\mathbf{E}\{\|x_{m_k}(y'_{k-1},\beta_k) - x_*\|_2^2\} \le 2^{-k}r_0^2 = r_k^2,$$

and (58) follows. Now (59) is an immediate consequence of (58) and (62). □

Proof of the theorem. **1^o.** Let us start with the situation where no asymptotic stage takes place. Because we have assumed that N is large enough so that at least one preliminary stage took place this can only happen when either $m_0K \geq \frac{N}{2}$ or $m_1 \geq \frac{N}{2}$. Due to $m_0 > 1$, by (56) we have in the first case:

$$\mathbf{E}\{\|y_K - x_*\|^2\} \leq R_K^2 := 2^{-K}R_0^2 + \frac{32\varsigma_*^2\bar{s}\varkappa'}{\underline{\kappa}\nu\varkappa} \leq 2^{-K+1}R_0^2 \leq R_0^2\exp\left\{-\frac{cN\underline{\kappa}}{\Theta\varkappa\bar{s}\nu}\right\}$$

for some absolute $c > 0$. Furthermore, due to (55) we also have in this case

$$\mathbf{E}\{g(\widehat{x}_{m_0}(y_{K-1},\beta))\} - g_* \leq 2^{-K-4}\frac{\underline{\kappa}R_0^2}{\bar{s}} + \frac{2\varkappa'\varsigma_*^2}{\varkappa\nu} \leq 2^{-K-3}\frac{\underline{\kappa}R_0^2}{\bar{s}} \leq \frac{\underline{\kappa}R_0^2}{\bar{s}}\exp\left\{-\frac{cN\underline{\kappa}}{\Theta\varkappa\bar{s}\nu}\right\}.$$

Next, $m_1 \geq \frac{N}{2}$ implies that

$$\frac{\bar{s}}{\underline{\kappa}} \geq \frac{cN}{\Theta\nu\varkappa} \tag{63}$$

for some absolute constant c, so that approximate solution y_K at the end of the preliminary phase satisfies (cf. (56))

$$\mathbf{E}\{\|\widehat{y} - x_*\|^2\} \leq C\frac{\varsigma_*^2\bar{s}\varkappa'}{\underline{\kappa}\nu\varkappa} \leq C\frac{\Theta\varkappa'\varsigma_*^2\bar{s}^2}{\underline{\kappa}^2N}.$$

Same as above, using (56) and (63) we conclude that in this case

$$\mathbf{E}\{g(\widehat{x})\} - g_* \leq C\frac{\varkappa'\varsigma_*^2}{\varkappa\nu} \leq C\frac{\Theta\varkappa'\varsigma_*^2\bar{s}}{\underline{\kappa}N}.$$

2^o. Now, let us suppose that at least one stage of the asymptotic phase was completed. Applying the bound (58) of Proposition A.2 we have $\mathbf{E}\{\|y'_K - x_*\|^2\} \leq r_0^2$. When $M < N/2$, same as above, we have

$$\mathbf{E}\{\|\widehat{y}_N - x_*\|^2\} \leq r_0^2 \leq R_0^2\exp\left\{-\frac{cN\underline{\kappa}}{\Theta\varkappa\bar{s}\nu}\right\}$$

and

$$\mathbf{E}\{g(\widehat{x}_{m_{K'}}(y_{K'-1},\beta))\} - g_* \leq \mathbf{E}\{g(\widehat{x})\} - g_* \leq \frac{\underline{\kappa}R_0^2}{\bar{s}}\exp\left\{-\frac{cN\underline{\kappa}}{\Theta\varkappa\bar{s}\nu}\right\}. \tag{64}$$

When $M \geq N/2$, since $m_k \leq C\bar{m}_k$ where $\bar{m}_k = 512\frac{\bar{s}\Theta\nu\varkappa}{\underline{\kappa}}2^k$ we have

$$\frac{N}{2} \leq C\sum_{k=1}^{K'}\bar{m}_k \leq C2^{K'+1}\bar{m}_1 \leq C2^{K'}\frac{\bar{s}\Theta\nu\varkappa}{\underline{\kappa}}.$$

We conclude that $2^{-K'} \leq C\frac{\bar{s}\Theta\nu\varkappa}{\underline{\kappa}N}$ so that

$$\mathbf{E}\{\|\widehat{y}_N - x_*\|^2\} = \mathbf{E}\{\|\widehat{y}_{K'} - x_*\|^2\} \leq 2^{-K'}r_0^2 \leq C\frac{\Theta\varkappa'\varsigma_*^2\bar{s}^2}{\underline{\kappa}^2N}.$$

Finally, by (59),

$$\mathbf{E}\{g(\widehat{x}_{m_{K'}}(y_{K'-1},\beta))\} - g_* \leq 2^{-K'+2}\frac{\varsigma_*^2\varkappa'}{\varkappa\nu} \leq C\frac{\varsigma_*^2\bar{s}\varkappa'\Theta}{\underline{\kappa}N};$$

together with (64) this implies (16). □

A.3 Proof of Theorem 2.2

1^o. By the Chebyshev inequality,

$$\forall \ell \quad \mathrm{Prob}\{\|\widehat{x}_M^{(\ell)} - x_*\|_2 \geq 2\theta_M\} \leq \tfrac{1}{4}; \tag{65}$$

applying [42, Theorem 3.1] we conclude that

$$\mathrm{Prob}\{\|\widehat{x}_{N,1-\epsilon} - x_*\|_2 \geq 2C_\alpha\theta_M\} \leq e^{-L\psi(\alpha,\frac{1}{4})}$$

where

$$\psi(\alpha,\beta) = (1-\alpha)\ln\frac{1-\alpha}{1-\beta} + \alpha\ln\frac{\alpha}{\beta} \tag{66}$$

and $C_\alpha = \frac{1-\alpha}{\sqrt{1-2\alpha}}$. When choosing $\alpha = \frac{\sqrt{3}}{2+\sqrt{3}}$ which corresponds to $C_\alpha = 2$ we obtain $\psi(\alpha,\frac{1}{4}) = 0.1070... > 0.1$ so that

$$\mathrm{Prob}\{\|\widehat{x}_{N,1-\epsilon} - x_*\|_2 \geq 4\theta_M\} \leq \epsilon$$

if $L \geq 10\ln[1/\epsilon]$. When combining this result with that of Lemma A.1 we arrive at the theorem statement for solutions $\widehat{x}_{N,1-\epsilon}$ and $\widehat{y}_{N,1-\epsilon}$.

2^o. The corresponding result for $\widehat{x}'_{N,1-\epsilon}$ and its "sparsification" $\widehat{y}'_{N,1-\epsilon}$ is due to the following simple statement.

Proposition A.3 *Let $0 < \alpha < \frac{1}{2}$, $|\cdot|$ be a norm on E, $z \in E$, and let z_ℓ, $\ell = 1,...,L$ be independent and satisfy*

$$\mathrm{Prob}\{|z_\ell - z| \geq \delta\} \leq \beta$$

for some $\delta > 0$ and $\beta < \alpha$. Then for $\widehat{z}$,

$$\widehat{z} \in \operatorname*{Argmin}_{u\in\{z_1,...,z_L\}} \sum_{\ell=1}^{L} |u - z_\ell|, \tag{67}$$

it holds

$$\mathrm{Prob}\{|\widehat{z} - z| \geq C'_\alpha\delta\} \leq e^{-L\psi(\alpha,\beta)}$$

with $C'_\alpha = \frac{2+\alpha}{1-2\alpha}$.

Proof. W.l.o.g. we may put $\delta = 1$ and $z = 0$. Proof of the proposition follows that of [42, Theorem 3.1] with Lemma 2.1 of [42] replaced with the following result.

Lemma A.2 *Let $z_1,...,z_L \in E$, and let $\widehat{z}$ be an optimal solution to (67). Let $0 < \alpha < \frac{1}{2}$, and let $|\widehat{z}| \geq C'_\alpha$. Then there exists a subset I of $\{1,...,L\}$ of cardinality $\mathrm{card}I > \alpha L$ such that for all $\ell \in I$ $|z_\ell| > 1$.*

Proof of the lemma. Let us assume that $|z_\ell| \leq 1$, $\ell = 1,...,\bar{L}$ for $\bar{L} \geq (1-\alpha)L$. Then

$$\begin{array}{rcl}
\sum_{\ell=1}^L |z_\ell - \widehat{z}| & = & \sum_{\ell\leq\bar{L}} |z_\ell - \widehat{z}| + \sum_{\ell>\bar{L}} |z_\ell - \widehat{z}| \geq \bar{L}(C_\alpha - 1) + \sum_{\ell>\bar{L}}[|z_\ell| - C_\alpha] \\
& \geq & \sum_{\ell\leq\bar{L}} |z_\ell| + \bar{L}(C_\alpha - 2) + \sum_{\ell>\bar{L}} |z_\ell| - (L-\bar{L})C_\alpha \\
& \geq & \sum_{\ell=1}^L |z_\ell| + \bar{L}(C_\alpha - 2) - (L-\bar{L})C_\alpha \\
& \geq & \sum_{\ell=1}^L |z_\ell - z_1| + \bar{L}(2C_\alpha - 2) - LC_\alpha + L - 1 > \sum_{\ell=2}^L |z_\ell - z_1|
\end{array}$$

for $\bar{L} > \frac{LC_\alpha + L - 1}{2(C_\alpha - 1)}$. We conclude that $1 - \alpha \leq \frac{C_\alpha + 1}{2(C_\alpha - 1)}$, same as $C_\alpha \leq \frac{2+\alpha}{1-2\alpha}$. □

For instance, when choosing $\alpha = 1/6$ with $C_\alpha = 13/4$, and β such that $C_\alpha/\sqrt{\beta} = 10$ we obtain $\psi(\alpha,\beta) = 0.0171...$ so that for $L = \lceil 58.46\ln[1/\epsilon]\rceil$ we have $L\psi(\alpha,\beta) \geq \ln[1/\epsilon]$. Because

$$\mathrm{Prob}\left\{\|\widehat{x}_M^{(\ell)} - x_*\|_2 \geq \frac{\theta_M}{\sqrt{\beta}}\right\} \leq \beta, \quad \ell = 1,...,L,$$

by Lemma A.2 we conclude that

$$\mathrm{Prob}\left\{\|\widehat{x}'_{1-\epsilon,N} - x_*\|_2 \geq 10\theta_M\right\} \leq \epsilon,$$

implying statement of the theorem for $\widehat{x}'_{1-\epsilon,N}$ and $\widehat{y}'_{1-\epsilon,N}$.

3^o. The proof of the claim for solutions $\widehat{x}''_{1-\epsilon,N}$ and $\widehat{y}''_{1-\epsilon,N}$ follows the lines of that of [25, Theorem 4]. We reproduce it here (with improved parameters of the procedure) to meet the needs of the proof of Theorem 2.3.

Let us denote $I(\tau_M)$ the subset of $\{1,...,L\} \cup \emptyset$ such that $g(\widehat{x}_M^{(i)}) - g_* \leq 2\tau_M$ and thus $\|\widehat{x}_M^{(i)} - x_*\|_2 \leq 2\theta_M$ for $i \in I(\tau_M)$. Assuming the latter set is nonempty we have for all $i,j \in I(\tau_M)$ $\|\widehat{x}_M^{(i)} - \widehat{x}_M^{(j)}\|_2 \leq 4\theta_M$. On the other hand, using (65) and independence of $\widehat{x}_M^{(i)}$ we conclude that (cf. e.g., [40, Lemma 23])

$$\mathrm{Prob}\left\{|I| \geq \rceil L/2\lceil\right\} \geq \mathrm{Prob}\left\{B(L,\tfrac{1}{4}) \geq \rceil L/2\lceil\right\} \geq 1 - \exp\left\{-L\psi\left(\frac{\lfloor L/2\rfloor}{L}, \frac{1}{4}\right)\right\}$$

where $\lfloor a\rfloor = \lceil a\rceil - 1$ is the largest integer strictly less than a, $B(N,p)$ is a (N,p)-binomial random variable and $\psi(\cdot,\cdot)$ is as in (66). When $\varepsilon \leq \frac{1}{4}$ and $L = \lceil 12.05\ln[1/\varepsilon]\rceil \geq 16$ we have

$$\mathrm{Prob}\{|I| \geq \rceil L/2\lceil\} \geq 1 - e^{-L\psi(\frac{7}{16},\frac{1}{4})} \geq 1 - e^{-0.083L} \geq 1 - \varepsilon.$$

Therefore, if we denote $\overline{\Omega}_\varepsilon$ a subset of Ω^N such that $|I(\tau_M)| > L/2$ for $\omega^N \in \overline{\Omega}_\epsilon$ we have $P\{\overline{\Omega}_\varepsilon\} \geq 1 - \varepsilon$. Let now $\omega^N \in \overline{\Omega}_\varepsilon$ be fixed. Observe that the optimal value $\widehat{r} = \widehat{r}^{\,i}_{\rceil L/2\lceil}$ of (21) satisfies $\widehat{r} \leq 4\theta_M$, and that among $\rceil L/2\lceil$ closest to $\widehat{x}''_{N,1-\epsilon}$ points there is at least one, let it be $\widehat{x}_M^{(\bar{i})}$ satisfying $g(\widehat{x}_M^{(\bar{i})}) - g_* \leq 2\tau_M$ and $\|\widehat{x}_M^{(\bar{i})} - x_*\|_2 \leq 2\theta_M$. We conclude that whenever $\omega^N \in \overline{\Omega}$ one has

$$\|\widehat{x}''_{N,1-\epsilon} - x_*\|_2 \leq \|\widehat{x}''_{N,1-\epsilon} - \widehat{x}_M^{(\bar{i})}\|_2 + \|\widehat{x}_M^{(\bar{i})} - x_*\|_2 \leq 4\theta_M + 2\theta_M \leq 6\theta_M,$$

implying that

$$\mathrm{Prob}\{\|\widehat{x}''_{N,1-\epsilon} - x_*\|_2 \geq 6\theta_M\} \leq \varepsilon$$

whenever $L \geq \lceil 12.05\ln[1/\varepsilon]\rceil$. □

A.4 Proof of Theorem 2.3

The proof of the theorem relies on the following statement which may be of independent interest.

Proposition A.4 *Let $U : [0,1] \times \Omega \to \mathbf{R}$ be continuously differentiable and such that $u(t) = \mathbf{E}\{U(t,\omega)\}$ is finite for all $t \in [0,1]$, convex and differentiable with Lipschitz-continuous gradient:*

$$|u'(t') - u'(t)|_* \leq \mathcal{M}|t - t'|, \qquad \forall\, t,t' \in [0,1].$$

In the situation in question, let $\varepsilon \in (0,\frac{1}{4}]$, $J \geq \left\lceil 7\ln[2/\varepsilon]\right\rceil$, and $t_i = \frac{2i-1}{2m}$, $i = 1,...,m$. Consider the estimate

$$\widehat{v} = \underset{j}{\mathrm{median}}[\widehat{v}^j], \quad \widehat{v}^j = \frac{1}{m}\sum_{i=1}^m U'(t_i,\omega_i^j) \quad j = 1,...,J$$

of the difference $v = u(1) - u(0)$ *using* $M = mJ$ *independent realizations* ω_i^j, $i = 1, ..., m$, $j = 1, ..., L$. *Then*

$$\mathrm{Prob}\{|\widehat{v} - v| \geq \rho\} \leq \varepsilon \tag{68}$$

where

$$\rho = \frac{1}{4m}\left[\sqrt{2\mathcal{M}(u(1) - u_*)} + \sqrt{2\mathcal{M}(u(0) - u_*)}\right] + \frac{2}{m}\sqrt{\sum_{i=1}^{m} \mathbf{E}\left\{[\zeta^1(t_i)]^2\right\}},$$

(here and below, $\zeta^j(t_i) = U'(t_i, \omega_i^j) - u'(t_i)$ *and* $u_* = \min_{0 \leq t \leq 1} u(t)$*).*

In particular, if for $\mu \geq \mathcal{M}$

$$\mathbf{E}\{[\zeta^1(t)]^2\} \leq \mu(u(t) - u_*) + \varsigma^2 \tag{69}$$

then

$$\mathrm{Prob}\{|\widehat{v} - v| \geq \bar{\rho}\} \leq \varepsilon \tag{70}$$

where

$$\bar{\rho} = 2\sqrt{\frac{\mu}{m}}\left[\sqrt{u(1) - u_*} + \sqrt{u(0) - u_*}\right] + \frac{2\varsigma}{\sqrt{m}}.$$

We postpone the proof of the proposition to the end of this section.

1^o. Let $\omega^N \in \overline{\Omega}_{\epsilon/2}$ defined as in 3^o of the proof of Theorem 2.2; we choose $L \geq \lceil 12.05 \ln[2/\varepsilon] \rceil$ so that $\mathrm{Prob}\{\overline{\Omega}_{\epsilon/2}\} \leq \epsilon/2$. We denote $\widehat{r}$ the optimal value of (21); recall that $\widehat{r} \leq 4\theta_M$. Then for any $i, j \in \widehat{I}$ we have

$$\|\widehat{x}_M^{(i)} - \widehat{x}_M^{(j)}\|_2 \leq 2\widehat{r} \leq 8\theta_M, \tag{71}$$

and for some $\bar{i} \in \widehat{I}$ we have

$$g(\widehat{x}_M^{(\bar{i})}) - g_* \leq 2\tau_M^2 \tag{72}$$

where τ_M and θ_M are defined in (18) and (19) respectively. W.l.o.g. we can assume that $\widehat{x}_M^{(\bar{i})}$ is the minimizer of $g(x)$ over $\widehat{x}_M^{(i)}$, $i \in \widehat{I}$.

Let us consider the aggregation procedure. From now on all probabilities are assumed to be computed with respect to the distribution P^K of the (second) sample ω^K, conditional to realization ω^N of the first sample (independent of ω^K). To alleviate notation we drop the corresponding "conditional indices."

2^o. Denote $\widehat{v}_{j\bar{i}} = \mathrm{median}_\ell[\widehat{v}_{j\bar{i}}^\ell]$. For $j \in \widehat{I}$, $j \neq \bar{i}$ let $x(t) = \widehat{x}_M^{(j)} + t\big(\widehat{x}_M^{(\bar{i})} - \widehat{x}_M^{(j)}\big)$. Note that $U(t, \omega) = G(x(t), \omega)$ and $u(t) = g(x(t))$ satisfy the premise of Proposition A.4 with $\mathcal{M} = r_{j\bar{i}}^2 \mathcal{L}_2$ where $r_{j\bar{i}} = \|\widehat{x}_M^{(\bar{i})} - \widehat{x}_M^{(j)}\|_2$, $\mu = \chi \mathcal{L}_2 r_{j\bar{i}}^2$, and $\varsigma^2 = \chi' \varsigma_*^2 r_{j\bar{i}}^2$. When applying the proposition with $\varepsilon = \epsilon/L$, $J = L'$, and $K = mL'$ we conclude that

$$\forall j \in \widehat{I},\, j \neq \bar{i} \quad \mathrm{Prob}\{|\widehat{v}_{j\bar{i}} - v_{j\bar{i}}| \geq \varrho_{j\bar{i}}\} \leq \frac{\epsilon}{L},$$

implying that

$$\mathrm{Prob}\{\max_{j \in \widehat{I}, j \neq \bar{i}} |\widehat{v}_{j\bar{i}} - v_{j\bar{i}}| \geq \varrho_{j\bar{i}}\} \leq \frac{\epsilon}{2} \tag{73}$$

where

$$\varrho_{ij} = 2r_{j\bar{i}}\sqrt{\frac{\mathcal{L}_2\chi}{m}}\left[\sqrt{g(\widehat{x}_M^{(i)}) - g_*} + \sqrt{g(\widehat{x}_M^{(j)}) - g_*}\right] + 2r_{j\bar{i}}\varsigma_*\sqrt{\frac{\chi'}{m}}.$$

Let now $\Omega'_{\epsilon/2} \subset \Omega^K$ such that for all

$$\max_{\bar{i}\neq j\in\widehat{I}} |\widehat{v}_{j\bar{i}} - v_{j\bar{i}}| \leq \varrho_{j\bar{i}}, \quad \forall \omega^K \in \Omega'_{\epsilon/2};$$

by (73) $\mathrm{Prob}\{\Omega'_{\epsilon/2}\} \geq 1 - \epsilon/2$.

3^o. Let us fix $\omega^K \in \Omega'_{\epsilon/2}$; our current objective is to show that in this case the set of admissible $\widehat{x}_M^{(i)}$'s is nonempty—it contains $\widehat{x}_M^{(\bar{i})}$—and, moreover, all admissible $\widehat{x}_M^{(j)}$'s satisfy the bound $g(\widehat{x}_M^{(j)}) \leq \gamma^2(r_{\bar{i}j})$ with $\gamma(r)$ defined as in (24).

Let $\alpha, \beta, \tau > 0$, and let $v(\gamma) = \gamma^2 - \tau^2 - 2[\alpha(\gamma+\tau)+\beta]$; then $v(\gamma) > 0$ for $\gamma \geq \sqrt{(2\alpha+\tau)^2+4\beta}$. Indeed, $v(\cdot)$ being nondecreasing for $\gamma \geq \alpha$, it suffices to verify the inequality for $\gamma = \sqrt{(2\alpha+\tau)^2+4\beta}$. Because

$$2\alpha + \tau + \beta/\alpha > \sqrt{(2\alpha+\tau)^2+4\beta}$$

we have

$$4\alpha^2 + 4\alpha\tau + 2\beta > 2\alpha\left(\sqrt{(2\alpha+\tau)^2+4\beta} + \tau\right),$$

and

$$v(\gamma) = [(2\alpha+\tau)^2+4\beta] - \tau^2 - 2\alpha\left(\sqrt{(2\alpha+\tau)^2+4\beta} + \tau\right) - 2\beta > 0.$$

Applying the above observation to $\alpha = 2r_{j\bar{i}}\sqrt{\frac{\mathcal{L}_2\chi}{m}}$, $\beta = 2r_{j\bar{i}}\varsigma_*\sqrt{\frac{\chi'}{m}}$, and $\tau = \tau_M$ we conclude that whenever $g(\widehat{x}_M^{(j)}) - g_* \geq \gamma^2(r_{j\bar{i}})$

$$v_{j\bar{i}} = g(\widehat{x}_M^{(\bar{i})}) - g(\widehat{x}_M^{(j)}) \leq \tau_M^2 - g(\widehat{x}_M^{(j)}) < -2\varrho_{j\bar{i}}. \tag{74}$$

Therefore, for $g(\widehat{x}_M^{(j)}) \geq \gamma^2(r_{j\bar{i}})$

$$\underset{\ell}{\mathrm{median}}[\widehat{v}_{j\bar{i}}^{\ell}] - \rho_{\bar{i}j} = [\underset{\ell}{\mathrm{median}}[\widehat{v}_{j\bar{i}}^{\ell}] - v_{j\bar{i}}] + v_{j\bar{i}} - \rho_{\bar{i}j} < \varrho_{j\bar{i}} - 2\varrho_{j\bar{i}} - \rho_{\bar{i}j} < 0 \ \ \forall \omega^K \in \Omega'_{\epsilon/2}.$$

Furthermore, for $g(\widehat{x}_M^{(j)}) - g_* < \gamma^2(r_{j\bar{i}})$ we have

$$\underset{\ell}{\mathrm{median}}[\widehat{v}_{j\bar{i}}^{\ell}] - \rho_{\bar{i}j} \leq \varrho_{\bar{i}j} - \rho_{\bar{i}j} < 0 \quad \forall \omega^K \in \Omega'_{\epsilon/2},$$

and we conclude that $\widehat{x}_M^{(\bar{i})}$ is admissible.

On the other hand, whenever $g(\widehat{x}_M^{(j)}) - g_* \geq \gamma^2(r_{j\bar{i}})$ we have $v_{\bar{i}j} > 2\varrho_{\bar{i}j}$ (cf. (74)), and

$$\underset{\ell}{\mathrm{median}}[\widehat{v}_{\bar{i}j}^{\ell}] - \rho_{j\bar{i}} = [\underset{\ell}{\mathrm{median}}[\widehat{v}_{j\bar{i}}^{\ell}] - v_{\bar{i}j}] + v_{\bar{i}j} - \rho_{\bar{i}j} > -\varrho_{\bar{i}j} + 2\varrho_{\bar{i}j} - \rho_{\bar{i}j} \geq 0 \ \ \forall \omega^K \in \Omega'_{\epsilon/2}.$$

We conclude that $\widehat{x}_M^{(j)}$ is not admissible if $g(\widehat{x}_M^{(j)}) \geq \gamma^2(r_{j\bar{i}})$ and $\omega^K \in \Omega'_{\epsilon/2}$.

4^o. Now we are done. So, assume that $[\omega^N, \omega^K] \in \overline{\Omega}_{\epsilon/2} \times \Omega'_{\epsilon/2}$ (what is the case with probability $\geq 1-\epsilon$). We have $r_{ij} \leq 8\theta_M$ for $i, j \in \widehat{I}$ by (71), and $g(x_M^{(\bar{i})}) \leq \tau_M^2$ for some admissible $\bar{i} \in \widehat{I}$ by (72). In this situation, all $\widehat{x}_M^{(j)}$ such that $g(\widehat{x}_M^{(j)}) - g_* \geq \gamma^2(r_{j\bar{i}})$, $j \in \widehat{I}$, are not admissible, implying that the suboptimality of the selected solution $\overline{x}_{N+K,1-\epsilon}$ is bounded with $\gamma^2(8\theta_M)$, thus

$$\mathrm{Risk}_{g,\epsilon}(\overline{x}_{N+K,1-\epsilon}|X) \leq \bar{\gamma}^2 = \gamma^2(8\theta_M).$$

The "in particular" part of the statement of the theorem can be verified by direct substitution of the corresponding values of m, θ_M, and τ_M into the expression for $\bar{\gamma}^2$. □

Proof of Proposition A.4. Let us denote

$$\bar{v} = \mathbf{E}\{\widehat{v}^j\} = \frac{1}{m}\sum_{i=1}^{m} u'(t_i);$$

we have

$$|\widehat{v} - v| \le |\widehat{v} - \bar{v}| + |\bar{v} - v|. \tag{75}$$

1^o. Note that

$$\widehat{v}^j - \bar{v} = \frac{1}{m}\sum_{i=1}^{m} U'(t_i, \omega_i^j) - u'(t_i) = \frac{1}{m}\sum_{i=1}^{m} \zeta^j(t_i),$$

and

$$\mathbf{E}\{(\widehat{v}^j - \bar{v})^2\} \le \frac{1}{m^2}\sum_{i=1}^{m} \mathbf{E}\{[\zeta^j(t_i)]^2\} =: \upsilon^2.$$

By the Chebyshev inequality, $\mathrm{Prob}\{|\widehat{v}^j - \bar{v}| \ge 2\upsilon\} \le \frac{1}{4}$, and

$$\begin{aligned}\mathrm{Prob}\{\operatorname*{median}_j[\widehat{v}^j] - \bar{v} \ge 2\upsilon\} &\le \mathrm{Prob}\Big\{\sum_j 1\{\widehat{v}^j - \bar{v} \ge 2\upsilon\} \ge J/2\Big\} \\ &\le \mathrm{Prob}\{B(J, \tfrac{1}{4}) \ge J/2\} \le e^{-J\psi(\frac{1}{2},\frac{1}{4})} \le e^{-0.1438J}\end{aligned}$$

where $\psi(\cdot,\cdot)$ is defined in (66). Because the same bound holds for $\mathrm{Prob}\{\mathrm{median}_j[\widehat{v}^j] - \bar{v} \le -2\upsilon\}$ we conclude that

$$\mathrm{Prob}\{|\widehat{v} - \bar{v}| \ge 2\upsilon\} = \mathrm{Prob}\{|\operatorname*{median}_j[\widehat{v}^j] - \bar{v}| \ge 2\upsilon\} \le 2e^{-J/7} \le \varepsilon \tag{76}$$

for $J \ge 7\ln(2/\varepsilon)$. Furthermore, if (69) holds we have

$$\mathbf{E}\{(\widehat{v}^j - \bar{v})^2\} \le \frac{1}{m^2}\sum_{i=1}^{m}[\mu(u(t_i) - g_*) + \varsigma^2] \le \frac{1}{2m}[(u(1) - u*) + (u(0) - u_*)] + \frac{\varsigma^2}{m} =: \bar{\upsilon}^2$$

implying (76) with υ replaced with $\bar{\upsilon}$:

$$\mathrm{Prob}\{|\widehat{v} - \bar{v}| \ge 2\bar{\upsilon}\} \le 2e^{-J/7} \le \varepsilon \tag{77}$$

2^o. Next, we bound the difference $\bar{v} - v$. Let $s_i = i/m$, $i = 0, ..., m$, and $r_i = u'(s_i) - u'(s_{i-1})$. Let us show that

$$v - \bar{v} \le \frac{1}{4m}\left[\sqrt{2\mathcal{M}(u(1) - u_*)} + \sqrt{2\mathcal{M}(u(0) - u_*)}\right].$$

Note that

$$\delta_i = \int_{s_{i-1}}^{s_i} [u'(s) - u'(t_i)]ds \le \tfrac{1}{4} r_i(s_i - s_{i-1}) = (4m)^{-1} r_i,$$

so that

$$v - \bar{v} \le \sum_{i=1}^{m} \delta_i \le (4m)^{-1}[u'(1) - u'(0)].$$

Let now $t_* \in [0,1]$ be a minimizer of u on $[0,1]$. Due to the smoothness and convexity of u we have

$$|u'(0) - u'(t_*)|^2 \le 2\mathcal{M}[u(0) - u_* + t_* u'(t_*)] \le 2\mathcal{M}[u(0) - u_*]$$

and

$$|u'(1) - u'(t_*)|^2 \le 2\mathcal{M}[u(1) - u_* - (1-t_*)u'(t_*)] \le 2\mathcal{M}[u(1) - u_*].$$

We conclude that

$$u'(1) - u'(0) \le u'(1) - u'(t_*) + u'(t_*) - u'(0) \le \sqrt{2\mathcal{M}[u(0) - u_*]} + \sqrt{2\mathcal{M}[u(1) - u_*]},$$

and

$$v - \bar{v} \le (4m)^{-1}[u'(1) - u'(0)] \le \frac{1}{4m}\left[\sqrt{2\mathcal{M}[u(0) - u_*]} + \sqrt{2\mathcal{M}[u(1) - u_*]}\right].$$

The proof of the corresponding bound for $\bar{v} - v$ is completely analogous, implying that

$$|v - \bar{v}| \le \frac{1}{4m}\left[\sqrt{2\mathcal{M}(u(1) - u_*)} + \sqrt{2\mathcal{M}(u(0) - u_*)}\right].$$

When substituting the latter bound and the bound (76) into (75) we obtain

$$\mathrm{Prob}\{|\widehat{v} - v| \ge 2\upsilon + \upsilon'\} \le \varepsilon$$

for $J \ge 7\ln(2/\varepsilon)$, what implies (68). When replacing (76) with (77) in the above derivation we obtain (70). □

A.5 Proofs for Section 3.2

The following statement is essentially well known:

Lemma A.3 *Let $\phi \in \mathbf{R}^{p\times q}$ with $q \le p$ for the sake of definiteness, be a random sub-Gaussian matrix $\phi \sim \mathrm{Sub}\mathcal{G}(0,S)$ implying that*

$$\forall x \in \mathbf{R}^{p\times q}, \quad \mathbf{E}\left\{e^{\langle x,\phi\rangle}\right\} \le e^{\frac{1}{2}\langle x, S(x)\rangle}. \tag{78}$$

Suppose that $S \preceq \bar{s}I$; then

$$\mathbf{E}\{\|\phi\|_*^2\} \le C\bar{s}(p+q) \quad \textit{and} \quad \mathbf{E}\{\|\phi\|_*^4\} \le C'\bar{s}^2(p+q)^2$$

where C and C' are absolute constants.

Proof of the lemma.

1^o. Let $u \in \mathbf{R}^q$ be such that $\|u\|_2 = 1$. Then the random vector $\zeta = \phi u \in \mathbf{R}^p$ is sub-Gaussian with $\zeta \sim \mathrm{Sub}\mathcal{G}(0,Q)$, that is for any $v \in \mathbf{R}^p$

$$\mathbf{E}\left\{e^{v^T\zeta}\right\} = \mathbf{E}\left\{e^{v^T\phi u}\right\} = \mathbf{E}\left\{e^{\langle uv^T,\phi\rangle}\right\} \le e^{\frac{1}{2}\langle uv^T, S(uv^T)\rangle} = e^{\frac{1}{2}v^TQv}$$

where $Q = Q^T \in \mathbf{R}^{p\times p}$. Note that

$$\max_{\|v\|_2=1} v^TQv = \max_{\|v\|_2=1} \langle uv^T, S(uv^T)\rangle \le \max_{\|w\|_2=1} \langle w, S(w)\rangle.$$

Therefore, we have $Q \preceq \bar{s}I$, and $\mathrm{Tr}(Q) \le \bar{s}p$.

2^o. Let $\Gamma = \{u \in \mathbf{R}^q : \|u\|_2 = 1\}$, and let $\mathcal{D}_\epsilon$ be a minimal ϵ-net, w.r.t. $\|\cdot\|_2$, in Γ, and let $\mathcal{N}_\epsilon$ be the cardinality of $\mathcal{D}_\epsilon$. We claim that

$$\left\{u^T\phi^T\phi u \le \upsilon \ \forall u \in \mathcal{D}_\epsilon\right\} \ \Rightarrow \ \left\{\|\phi^T\phi\|_* \le (1-2\epsilon)^{-1}\upsilon\right\}. \tag{79}$$

Indeed, let the premise in (79) hold true; $\phi^T\phi$ is symmetric, so let $\bar{v} \in \Gamma$ be such that $\bar{v}^T\phi^T\phi\bar{v} = \|\phi^T\phi\|_*$. There exists $u \in \mathcal{D}_\epsilon$ such that $\|\bar{v} - u\|_2 \le \epsilon$, whence

$$\|\phi^T\phi\|_* = |\bar{v}^T\phi^T\phi\bar{v}| \le 2\|\phi^T\phi\|_*\|\bar{v}-u\|_2 + |u^T\phi^T\phi u| \le 2\|\phi^T\phi\|_*\epsilon + \upsilon$$

(note that the quadratic form z^TQz is Lipschitz continuous on Γ, with constant $2\|Q\|_*$ w.r.t. $\|\cdot\|_2$), whence $\|\phi^T\phi\|_* \le (1-2\epsilon)^{-1}\upsilon$.

3^0. We can straightforwardly build an ϵ-net $\mathcal{D}'$ in Γ in such a way that the $\|\cdot\|_2$-distance between every two distinct points of the net is $> \epsilon$, so that the balls $B_v = \{z \in \mathbf{R}^p : \|z - v\|_2 \le \epsilon/2\}$ with $v \in \mathcal{D}'$ are mutually disjoint. Since the union of these balls belongs to $B = \{z \in \mathbf{R}^q : \|z\|_2 \le 1 + \epsilon/2\}$, we get $\mathrm{Card}(\mathcal{D}')(\epsilon/2)^q \le (1+\epsilon/2)^q$, that is, $\mathcal{N}_\epsilon \le \mathrm{Card}(\mathcal{D}') \le (1+2/\epsilon)^q$.

Now we need the following well-known result (we present its proof at the end of this section for the sake of completeness).

Lemma A.4 *Let $\zeta \sim \mathrm{Sub}\mathcal{G}(0, Q)$ be a sub-Gaussian random vector in $\mathbf{R}^n$, i.e.*

$$\forall t \in \mathbf{R}^n \quad \mathbf{E}\left\{e^{t^T\zeta}\right\} \le e^{\frac{1}{2}t^TQt} \tag{80}$$

where $Q = Q^T \in \mathbf{R}^{n\times n}$. Then for all $x \ge 0$

$$\mathrm{Prob}\{\|\zeta\|_2^2 \ge \mathrm{Tr}(Q) + 2\sqrt{xv} + 2x\bar{q}\} \le e^{-x} \tag{81}$$

where $\bar{q} = \max_i \sigma_i(Q)$ is the principal eigenvalue of Q and $v = \|Q\|_2^2 = \sum_i \sigma_i^2(Q)$ is the squared Frobenius norm of Q. Thus, for any $\alpha > 0$

$$\mathrm{Prob}\{\|\zeta\|_2^2 \ge \mathrm{Tr}(Q)(1+\alpha^{-1}) + (2+\alpha)x\bar{q}\} \le e^{-x}. \tag{82}$$

Utilizing (82) with $\alpha = 1$ we conclude that $\forall u \in \Gamma$ the random vector $\zeta = \phi u$ satisfies

$$\mathrm{Prob}\{\|\zeta\|_2^2 \ge 2\bar{s}p + 3\bar{s}x\} \le e^{-x}. \tag{83}$$

Let us set $\epsilon = \frac{1}{4}$; utilizing (83), we conclude that the probability of violating the premise in (79) with $\upsilon = 2\bar{s}p + 3\bar{s}x$ does not exceed $\exp\{-x + q\ln[1+2\epsilon^{-1}]\} = \exp\{-x + q\ln 9\}$, so that

$$\mathrm{Prob}\left\{\|\phi^T\phi\|_* \ge 2\bar{s}(2p+3x)\right\} \le \exp\{-x + q\ln 9\}.$$

Now we are done: recall that

$$\begin{aligned}
\mathbf{E}\{\|\phi\|_*^4\} &= \mathbf{E}\{\|\phi^T\phi\|_*^2\} = 2\int_0^\infty \mathrm{Prob}\{\|\phi^T\phi\|_* \ge u\}u\,du \\
&\le 2\int_0^\infty u\min\left\{\exp\left\{\frac{4\bar{s}p-u}{6\bar{s}} + q\ln 9\right\}, 1\right\}du \\
&\le 2\int_0^{\bar{s}(4p+6q\ln 9)} u\,du + 2\int_{\bar{s}(4p+6q\ln 9)}^\infty u\exp\left\{\frac{4\bar{s}p-u}{6\bar{s}} + q\ln 9\right\}du \\
&\le \bar{s}^2(4p+6q\ln 9)^2 + 12\bar{s}^2(4p+6q\ln 9) + 72\bar{s}^2 \le C'\bar{s}^2(p+q)^2.
\end{aligned}$$

Similarly we get $\mathbf{E}\{\|\phi\|_*^2\} \le C\bar{s}(p+q)$ for an appropriate C.

4^o. Let us now prove Lemma A.4.

Note that for $t < 1/(2\bar{s})$ and $\eta \in \mathbf{R}^n$, $\eta \sim \mathcal{N}(0, I)$ independent of ζ we have by (78)

$$\begin{aligned}\mathbf{E}\{e^{t\langle\zeta,\zeta\rangle}\} &= \mathbf{E}\left\{\mathbf{E}_\eta\{e^{\sqrt{2t}\langle\zeta,\eta\rangle}\}\right\} = \mathbf{E}_\eta\left\{\mathbf{E}\{e^{\sqrt{2t}\langle\zeta,\eta\rangle}\}\right\} \le \mathbf{E}_\eta\{e^{t\langle\eta,S\eta\rangle}\} = \mathbf{E}_\eta\{e^{t\langle\eta,D\eta\rangle}\} \\ &= \prod_i \mathbf{E}_{\eta_i}\{e^{t\eta_i^2 s_i}\} = \prod_i (1-2ts_i)^{-1/2}\end{aligned}$$

where $D = \mathrm{Diag}(s_i)$ is the diagonal matrix of eigenvalues. Recall that one has, cf. [6, Lemma 8],

$$-\tfrac{1}{2}\ln(1-2ts_i) - ts_i \le \frac{t^2 s_i^2}{1-2ts_i} \le \frac{t^2 s_i^2}{1-2t\bar{s}}$$

for $t < 1/(2\bar{s})$. On the other hand, $\forall t < 1/(2\bar{s})$

$$\begin{aligned}\mathrm{Prob}\{\|\zeta\|_2^2 - \mathrm{Tr}(S) \ge u\} &\le \mathbf{E}\Big\{\exp\Big\{t\big[\|\zeta\|_2^2 - \sum_i s_i - u\big]\Big\}\Big\} \\ &\le \exp\Big\{-tu + \frac{t^2}{1-2t\bar{s}}\sum_i s_i^2\Big\} = \exp\Big\{-tu + \frac{t^2 v}{1-2t\bar{s}}\Big\}.\end{aligned}$$

When choosing $t = \frac{\sqrt{x}}{v+2\bar{s}\sqrt{x}}$ $\left(< \frac{1}{2\bar{s}}\right)$ and $u = 2\sqrt{xv} + 2x\bar{s}$ we obtain

$$\mathrm{Prob}\{\|\zeta\|_2^2 \ge \mathrm{Tr}(S) + 2\sqrt{xv} + 2x\bar{s}\} \le e^{-x}$$

which is (81). Because $v \le \mathrm{Tr}(S)\bar{s}$ the latter bound also implies (82). □

References

[1] B. Adcock, A. C. Hansen, C. Poon, and B. Roman. Breaking the coherence barrier: A new theory for compressed sensing. In *Forum of Mathematics, Sigma*, volume 5. Cambridge University Press, 2017.

[2] A. Agarwal, S. Negahban, and M. J. Wainwright. Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions. In *Advances in Neural Information Processing Systems*, pages 1538–1546, 2012.

[3] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[4] A. Bietti and J. Mairal. Stochastic optimization with variance reduction for infinite datasets with finite sum structure. In *Advances in Neural Information Processing Systems*, pages 1623–1633, 2017.

[5] J. Bigot, C. Boyer, and P. Weiss. An analysis of block sampling strategies in compressed sensing. *IEEE transactions on information theory*, 62(4):2125–2139, 2016.

[6] L. Birgé, P. Massart, et al. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

[7] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

[8] C. Boyer, J. Bigot, and P. Weiss. Compressed sensing with structured sparsity and structured acquisition. *Applied and Computational Harmonic Analysis*, 46(2):312–350, 2019.

[9] E. Candes. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 1433–1452. Madrid, August 22-30, Spain, 2006.

[10] E. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus de l'Académie des Sciences, Mathématique*, 346(9-10):589–592, 2008.

[11] E. Candes and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. to appear. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[12] E. Candes, T. Tao, et al. The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351, 2007.

[13] E. J. Candes and Y. Plan. A probabilistic and ripless theory of compressed sensing. *IEEE transactions on information theory*, 57(11):7235–7254, 2011.

[14] E. J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[15] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[16] E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

[17] E. J. Candès, Y. Plan, et al. Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

[18] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[19] A. Dalalyan and P. Thompson. Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's m-estimator. In *Advances in Neural Information Processing Systems*, pages 13188–13198, 2019.

[20] T. Eltoft, T. Kim, and T.-W. Lee. On the multivariate laplace distribution. *IEEE Signal Processing Letters*, 13(5):300–303, 2006.

[21] M. Fazel, E. Candes, B. Recht, and P. Parrilo. Compressed sensing and robust recovery of low rank matrices. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1043–1047. IEEE, 2008.

[22] R. Foygel Barber and W. Ha. Gradient descent with non-convex constraints: local concavity determines convergence. *Information and Inference: A Journal of the IMA*, 7(4):755–806, 03 2018.

[23] P. Gaillard and O. Wintenberger. Sparse accelerated exponential weights. In *20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017. arXiv preprint arXiv:1610.05022.

[24] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

[25] D. Hsu and S. Sabato. Heavy-tailed regression with a generalized median-of-means. In *International Conference on Machine Learning*, pages 37–45, 2014.

[26] P. Jain, A. Tewari, and P. Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

[27] A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*, pages 121–148. MIT Press, 2011.

[28] A. Juditsky and A. Nemirovski. Accuracy guarantees for ℓ_1-recovery. *IEEE Transactions on Information Theory*, 57(12):7818–7839, 2011.

[29] A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, I: general purpose methods. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*. MIT Press Cambridge, 2011.

[30] A. Juditsky and Y. Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80, 2014.

[31] A. Juditsky, A. Nazin, A. Tsybakov, and N. Vayatis. Generalization error bounds for aggregation by mirror descent with averaging. In *Advances in neural information processing systems*, pages 603–610, 2006.

[32] A. Juditsky, F. K. Karzan, and A. Nemirovski. On a unified view of nullspace-type conditions for recoveries associated with general sparsity structures. *Linear Algebra and its Applications*, 441:124–151, 2014.

[33] V. Koltchinskii, K. Lounici, A. B. Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[34] S. Kotz and S. Nadarajah. *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.

[35] A. Kulunchakov. *Stochastic optimization for large-scale machine learning: variance reduction and acceleration*. PhD thesis, Université Grenoble Alpes, 2020. http://www.theses.fr/s192251.

[36] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

[37] G. Lecué and M. Lerasle. Robust machine learning by median-of-means: theory and practice. 2017. arXiv preprint arXiv:1711.10306.

[38] G. Lecué, S. Mendelson, et al. Regularization and the small-ball method i: sparse recovery. *The Annals of Statistics*, 46(2):611–641, 2018.

[39] S. Lee, J. Ha, O. Na, and S. Na. The cusum test for parameter change in time series models. *Scandinavian Journal of Statistics*, 30(4):781–796, 2003.

[40] M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. 2011. arXiv preprint arXiv:1112.3914.

[41] H. Liu and R. Foygel Barber. Between hard and soft thresholding: optimal iterative thresholding algorithms. *Information and Inference: A Journal of the IMA*, 9(4):899–933, 2020.

[42] S. Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4): 2308–2335, 2015.

[43] I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1-2):69–107, 2019.

[44] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

[45] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[46] A. S. Nemirovski and D. B. Yudin. *Complexity of problems and effectiveness of methods of optimization(Russian book)*. Nauka, Moscow, 1979. Translated as *Problem complexity and method efficiency in optimization,* J. Wiley & Sons, New York 1983.

[47] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

[48] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

[49] Y. Nesterov and A. Nemirovski. On first-order algorithms for ℓ_1/nuclear norm minimization. *Acta Numerica*, 22:509–575, 2013.

[50] N. Nguyen, D. Needell, and T. Woolf. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Transactions on Information Theory*, 63(11): 6869–6895, 2017.

[51] W. Ploberger and W. Krämer. The cusum test with ols residuals. *Econometrica: Journal of the Econometric Society*, pages 271–285, 1992.

[52] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug):2241–2259, 2010.

[53] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[54] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. In *Conference on Learning Theory, Workshop and Conference Proceedings*, volume 23, pages 10.1–10.28, 2012.

[55] S. Shalev-Shwartz and A. Tewari. Stochastic methods for l1-regularized loss minimization. *Journal of Machine Learning Research*, 12(Jun):1865–1892, 2011.

[56] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, pages 2199–2207, 2010.

[57] S. Van De Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.